THE UNIVERSITY *of* EDINBURGH

Statistical modelling of masked gene regulatory pathway changes across microarray studies of interferon gamma activated macrophages

Thorsten Forster

Doctor of Philosophy

The University of Edinburgh

2013

Declaration

I hereby declare that this thesis is of my own composition, and that it contains no material previously submitted for the award of any other degree. The work reported in this thesis has been executed by myself, except where due acknowledgement is made in the text.

Thorsten Forster

Acknowledgements

This will be lengthy, as I'm old and this part-time thesis has taken many a year to complete.

This research was in part supported by BBSRC (BB/J019283/1), Wellcome Trust (086696/Z/08/Z) and MRC (G0701291, G0701289).

I would like to thank my principal supervisor Prof Peter Ghazal for chaperoning me for many years through the vast and intricate maze that is host-pathogen biology, making high-volume high-throughput biological data my analytical playground, but also for encouragement and support in setting out on my own research at a more "mature" age. Without Peter endorsing the Possible, I could easily have followed the Likely and stayed within my presumed limitations.

Much gratitude also goes to Prof Chris Glasbey, who provided essential (probably statistically significant?) pointers and questions at those branching points in my research where a particular path needed to be hacked through the forest of estimates, statistics, metrics, models and parameters.

Thanks to my committee of Dr Guido Sanguinetti, Dr Anton Enright and Prof Richard Mott for making the viva experience a pleasant and productive one. Equations, attributions, punctuation, they made it better. This is not always a given.

I spent my early work years on statistical and computational subjects, and it is only within Peter's group that I was fully exposed to the complexities of biological research (I naively did use to think that a move from clinical to biological statistics would provide a wealth of less noisy data). It is only colleagues, students and collaborators that provided the essential subtitles and explained the plot lines of this foreign language film until I understood that the butler did it. Too many to list, but I'll highlight Dr Kevin Robertson and Dr Paul Dickinson, who have done much of the explaining all these years. In recent months, thanks to Dr Mathieu Blanc, Wayne Hsieh and Lu Hongjin for biological validation experiments.

Going back in time a bit further, a special thank you to Prof Robin Prescott, who got me into the country (specifically Edinburgh) as a summer student in 1996, provided my first professional employment a year later and did show me over several years with great patience and by great example how statistics and research studies really work (who knew that studying the stuff was only a meagre taster) and that this is a worthwhile profession to take up. Of course, Robin is therefore in not too minimal ways 'responsible' for me making my home in Edinburgh, with a lovely wife and family that now go along with it.

My mother deserves special thanks for her steady but unobtrusive determination that her sons would have choices in life. Unwisely, she even encouraged living abroad, but in the end the distance is only geographical. Thank you to my wife Eleanor who had to patiently wait for me to catch her up into doctorhood, proofread chapters, make cups of tea and just on the side provide the majority of my life's enjoyment. More good things to come.
My in-laws Margaret and Bill (whose presence, take on life and shared mediocrity at snooker will be missed after his recent passing) have over the last 9 years included me in a whole new and welcoming family; I could hardly have had a better backdrop for this research.

#Thank you

Abstract

Interferon gamma (IFN-γ) regulation of macrophages plays an essential role in innate immunity and pathogenicity of viral infections by directing large and small genome-wide changes in the transcriptional program of macrophages. Smaller changes at the transcriptional level are difficult to detect but can have profound biological effects, motivating the hypothesis of this thesis that responses of macrophages to immune activation by IFN-γ include small quantitative changes that are masked by noise but represent meaningful transcriptional systems in pathways against infection. To test this hypothesis, statistical meta-analysis of microarray studies is investigated as a tool to obtain the necessary increase in analysis sensitivity.

Three meta-analysis models (Effect size model, Rank Product model, Fisher's sum of logs) and three further modified versions were applied to a heterogeneous set of four microarray studies on the effect of IFN-γ on murine macrophages. Performance assessments include recovery of known biology and are followed by development of novel biological hypotheses through secondary analysis of meta-analysis outcomes in context of independent biological data sources. A separate network analysis of a microarray time course study investigate s if gene sets with coordinated time-dependent relationships overlap can also identify subtle IFN-γ related transcriptional changes in macrophages that match those identified through meta-analysis.

It was found that all meta-analysis models can identify biologically meaningful transcription at enhanced sensitivity levels, with slightly improved performance advantages for a non-parametric model (Rank Product meta-analysis). Meta-analysis yielded consistently regulated genes, hidden in individual microarray studies, related to sterol biosynthesis (*Stard3*, *Pgrmc1*, *Galnt6*, *Rab11a*, *Golga4*, *Lrp10*), implicated in cross-talk between type II and type I interferon or IL-10 signalling (*Tbk1*, *Ikbke*, *Clic4*, *Ptpre*, *Batf*), and circadian rhythm (*Csnk1e*). Further network analysis confirms that meta-analysis findings are highly concentrated in a distinct immune response cluster of co-expressed genes, and also identifies global expression modularisation in IFN-γ treated macrophages, pointing to *Trafd1* as a central anti-correlated node topologically linked to interactions with down-regulated sterol biosynthesis pathway members.

Outcomes from this thesis suggest that small transcriptional changes in IFN-γ activated macrophages can be detected by enhancing sensitivity through combination of multiple microarray studies. Together with use of bioinformatical resources, independent data sets and network analysis, further validation assigns a potential role for low or variable transcription genes in linking type II interferon signalling to type I and TLR signalling, as well as the sterol metabolic network.

Table of Contents

Declaration **2**
Acknowledgements **3**
Abstract **4**
Table of Contents **5**
List of Figures **8**
List of Tables **9**
List of Equations **10**
Abbreviations **11**
Chapter 1 ..... **13**
Introduction **13**
1.1 Macrophage biology and interferon signalling 13
1.2 Microarrays 20
1.2.1 Historical development *20*
1.2.2 Next-generation sequencing technology platforms *22*
1.2.3 Microarray technology overview *23*
1.2.4 Microarray processing and analysis *26*
1.2.5 Transcriptome *32*
1.3 Meta-analysis 33
1.3.1 Definition and background *33*
1.3.2 Meta-analysis key considerations *35*
1.4 Microarray meta-analysis 38
1.4.1 Literature summary *38*
1.4.2 Development of meta-analysis for microarray data *40*
1.4.3 Microarray meta-analysis workflow *45*
1.5 Thesis overview 46
Chapter 2 ..... **52**
Materials and Methods **52**
2.1 Microarray studies and data sets 52
2.1.1 Microarray data sources *52*
2.1.2 Identification of relevant microarray studies *53*
2.1.3 Combining microarray data sets *61*
2.2 Software tools and resources 62
2.2.1 Imputation of missing values *65*
2.3 Microarray target sample and array processing 66
2.4 Data quality control 67
2.5 Microarray data processing 70
2.6 Statistical analysis of individual microarray studies 72
2.7 IFN-γ pathway reference lists 75
2.7.1 NCBI gene reference lists *75*

2.7.2 Division of Pathway Medicine gene reference list 76
2.8 Reference sources for biological validation of meta-analysis results 77
2.8.1 Interferome 77
2.8.2 Microarray time course study (MITCH12) 78
2.8.3 Transcription factor binding site enrichment (oPOSSUM 3.0) 80

Chapter 3 81

Meta-analysis of the effect of IFN-γ on macrophage gene expression 81
3.1 Introduction 81
3.1.2 Data pre-processing 85
3.2 Meta-analysis models 86
3.2.1 Fisher's combined probability model 86
3.2.2 Rank-Product meta-analysis 88
3.3 Assessment methods for meta-analysis results 98
3.3.1 Pyramid plots and discovery counts 100
3.3.2 Integration-driven Discovery Rates (IDR) 102
3.3.3 Receiver-Operating-Characteristics (ROC) 104
3.3.4 Correspondence-at-Top (CAT) 105
3.4 Results 106
3.4.1 Discovery count results 106
3.4.2 IDR results 110
3.4.3 ROC results 115
3.4.4 CAT results 120
3.5 Discussion 123

Chapter 4 134

Imputation of missing values in microarray meta-analyses 134
4.1 Introduction to microarray missing value imputation 135
4.2 Imputation algorithms 139
4.3 Evaluation methodology 143
4.4 Results 144
4.5 Discussion 153

Chapter 5 160

Biological context of meta-analysis results 160
5.1 Introduction 160
5.2 IFN-γ signal transduction 162
5.3 Known interferon stimulated genes (ISGs) 165
5.4 Other evidence of ISGs 166
5.4.1 Experimental evidence for ISGs provided by DPM 166
5.4.2 Promoter predicted GAS binding site evidence for ISGs 167
5.5 Meta-analysis result gene sets 167
5.5.1 Meta-analysis results prior to aggregation 168
5.5.2 Consensus lists of meta-analysis results 170
5.6 Meta-analysis results in biological context 175
5.6.1 In the context of JAK-STAT pathway 175
5.6.2 In the context of known ISGs 176
5.6.3 New candidate ISGs based on existing annotation 180
5.6.4 New candidate ISGs based on time course data 182
5.6.5 Candidate ISGs based on other evidence 187
5.7 Master lists of identified candidate genes 188
5.8 Discussion 193

5.8.1 Discussion of issues relevant to result interpretation 193
5.8.2 Discussion of genes in biological context 196

Network analysis of an IFN-γ time course study 202
6.1 Macrophage activation and function 203
6.2 Microarray time course study 205
6.3 Network analysis introduction 205
6.4 Network analysis 206
6.4.1 Input data 206
6.4.2 Computation of adjacency matrices 207
6.4.3 Size reduction of adjacency matrices 207
6.4.4 Removal of redundant gene-gene correlations through PCIT 207
6.4.5 Basic visualisation of all remaining gene-gene associations 209
6.4.6 "Guilt-by-association" networks 209
6.5 Results 209
6.5.1 Size reduction of initial adjacency matrices has notably different effects on the IFN-γ time course 212
6.5.2 Gene cross-correlation maps show structural differences between the three time-courses 213
6.5.3 Cytoscape gene co-expression networks show structural differences between the three time courses as well as particular gene "hubs" 216
6.5.4 Cytoscape gene co-expression networks show alignment with meta-analysis results 217
6.5.5 Cytoscape gene co-expression networks may indicate different biological functionality. 220
6.5.6 Trafd1 (alias Fln29) 221
6.6 Discussion 228

Chapter 7 233

Discussion 233
7.1 Research conclusions 234
7.2 Implications of research outcomes 236
7.2 Limitations of research outcomes 241
7.3 Future research 245
7.4 Final thoughts 248

Appendix 250
A1- Protein network (STRING), down regulated genes 251
A2 - Protein network (STRING), up regulated genes 252
A3 – QQ plot for deciding on effect size model 253
A4 – ROC plots for NCBI-derived gene reference lists 254
A5 – Sensitivity and specificity summary 255
A6 – Reversed IDR plots, Fisher's meta-analysis 256
A7 – Reversed IDR plots, Effect Size meta-analysis 257
A8 – Reversed IDR plots, Rank Product meta-analysis 258
A9 – Reversed IDR plots, Rank Product meta-analysis, proportional weights 259
A10 – Reversed IDR plots, Rank Product meta-analysis, equal weights 260
A11 – Transcription factor enrichment of aggregated meta-analysis results 261
A12 – DPM gene reference list ("gold standard") 262
Appendix – Publications (co/authored) 264

Bibliography 269

List of Figures

Figure 1.1 Type II interferon signalling 1 ..... *19*
Figure 1.2 Affymetrix GeneChip processing *25*
Figure 1.3 Microarray processing and analysis *26*
Figure 1.4 Publications with "microarray" and "meta-analysis" in title *39*
Figure 1.5 Publications with "next" "generation" "sequencing" "analysis" in title *39*
Figure 1.6 Publications with "microarray" and "meta-analysis" as topic *40*
Figure 1.7 Microarray meta-analysis workflow *46*
Figure 2.1 Array quality control graphs *68*
Figure 2.2 MITCH12 study design *79*
Figure 3.1.1 Study selection, analysis and meta-analysis workflow *83*
Figure 3.1.2 Effect Size model overview *94*
Figure 3.2 Example of a standard forest plot *99*
Figure 3.3 Example pyramid plot *101*
Figure 3.4 Discoveries in individual studies and meta-analyses *109*
Figure 3.5 Overlap between individual studies and meta-analysis results *111*
Figure 3.6 IDR plots for meta-analysis-only discoveries *112*
Figure 3.7 ROC curves for meta-analysis compared to DPM reference *116*
Figure 3.8 ROC curves for study results compared to DPM reference *118*
Correspondence-At-Top assessment as used in this thesis is somewhat similar to ROC in that it uses the same biological reference list, but it highlights the number of ordered significant results that overlap with the reference list. Figure 3.9 compares this metric between the meta-analysis models. *120*
Figure 3.9 Correspondence-At-Top using DPM reference *120*
Figure 3.10 CAT for per-study results using DPM reference *122*
Figure 4.1 Missing gene expression values in microarray studies *137*
Figure 4.2 Imputation accuracy for individual arrays *145*
Figure 4.3 Recovery of differential expression values by array-wise linear regression imputation *148*
Figure 4.4 Recovery of differential expression values by KNN imputation *149*
Figure 4.5 Comparison of meta-analysis on missing or imputed data to original meta-analysis (missing values frequency 17%) *152*
Figure 4.6 Comparison of meta-analysis on missing or imputed data to original meta-analysis (missing value frequency 1.5%) *153*
Figure 5.1 Canonical IFN-γ activated JAK-STAT signalling *164*
Figure 5.2 Venn diagram defining consensus gene set *174*
Figure 5.3 Expression profiles of selected new identified ISGs *186*
Figure 6.1 Overview of PCIT algorithm *208*
Figure 6.2 Overview heatmap of clustered gene expression profiles *210*
Figure 6.3 Count of highly-expressed genes per time-point *211*
Figure 6.4 Gene cross-correlation maps *215*
Figure 6.5a IFN-γ treatment, gene co-expression network *218*
Figure 6.5b Virus-infection, gene co-expression network *219*
Figure 6.5c Combined virus+IFN-γ, gene co-expression network *220*
Figure 6.6 Trafd1 and Srebf2 expression profiles on IFN-γ treatment of macrophages *222*
Figure 6.7 Trafd1 expression in an independent time course study *224*
Figure 6.8 Trafd1 mRNA expression validation *225*
Figure 6.9 Trafd1 laboratory validation in siRNA knock-down *226*
Figure 6.10 Schematic overview of known Trafd1 function *227*
Figure 7.1 Genes of relevance to this thesis *243*

List of Tables

Table 2.1 Overview of all microarray studies used in thesis *58*
Table 2.2 Description of study 1 ..... *58*
Table 2.3 Description of study 2 ..... *59*
Table 2.4 Description of study 3 ..... *59*
Table 2.5 Description of study 4 ..... *60*
Table 2.6 Description of study 5 ..... *60*
Table 2.7 Description of study 6 ..... *61*
Table 2.8 Existing R/Bioconductor packages used *63*
Table 2.9 List of bioinformatics resources *64*
Table 2.10 Query-based reference list of known biological pathways *76*
Table 3.1 Per-study group sample sizes *85*
Table 5.1 Summary of meta-analysis-only discoveries *169*
Table 5.2 High-confidence gene lists *171*
Table 5.3 Meta-analysis results compared to Interferome.org ISGs *176*
Table 5.4 Meta-analysis results compared to in-silico ISGs *177*
Table 5.5 Transcription factors in meta-analysis results *178*
Table 5.6 New ISGs suggested by existing annotation *181*
Table 5.7 New ISGs suggested by IFN-γ microarray time course study *182*
Table 5.8 Master list of IFN-γ up regulated meta-analysis-only discoveries with supporting evidence *189*
Table 5.9 Master list of IFN-γ down regulated meta-analysis-only discoveries with supporting evidence *191*
Table 6.1 Adjacency matrix size reduction steps *213*
Table 6.2 Trafd1 correlations to sterol pathway genes *223*

List of Equations

Eq. 3.1 Fisher's combination of probabilities *87*
Eq. 3.2 Expression ratio Q between sample pairs *89*
Eq. 3.3 Rank of expression ratios *89*
Eq. 3.4 Rank Product *90*
Eq. 3.5 Rank Product permutations *90*
Eq. 3.6 Probability of observed Rank Product *90*
Eq. 3.7 Per-study weights for Rank Product *91*
Eq. 3.8 Proportional study weighting *91*
Eq. 3.9 Equal study weighting *91*
Eq. 3.10 Cohen's d (effect size) *95*
Eq. 3.11 Unbiased Cohen's d *95*
Eq. 3.12 Variance of unbiased Cohen's d *95*
Eq. 3.13 Cochran's Q *96*
Eq. 3.14 Per-study weights *96*
Eq. 3.15 Overall treatment effect *96*
Eq. 3.16 DerSimonian-Laird heterogeneity estimator *97*
Eq. 3.17 Weighted overall effect size *97*
Eq. 3.18 Variance of overall effect size *97*
Eq. 3.19 z-score for overall effect size *98*
Eq. 3.20 Integration-Driven Discovery gene status (IDD_g) *103*
Eq. 3.21 Integratio-Driven Discovery (IDD) *103*
Eq. 3.22 Integration-driven Discovery Rate (IDR) *103*

Abbreviations

ANOVA	Analysis Of Variance
AUC	Area Under Curve
BMDM	Bone marrow derived macrophages
BPCA	Bayesian Principal Components Analysis
CAT	Correspondance-At-Top
DAVID	Database for Annotation, Visualization and Integrated Discovery
DPM	Division of Pathway Medicine
ES	Effect Size
FDR	False Discovery Rate
FEM	Fixed Effect Model
FP	Fisher's combined Probabilities
GAS	Gamma Activated Sequence
GEO	Gene Expression Omnibus
GO	Gene Ontology
GPX-MEA	GPX-Macrophage Expression Atlas
IDD	Integration-Driven Discovery
IDR	Integration-driven Discovery Rate
IFN	Interferon
IFNAR	Interferon alpha receptor
IFNGR	Interferon gamma receptor
IPA	Ingenuity Pathway Analyzer
IRG	Interferon Regulated Gene
ISG	Interferon Stimulated Gene
ISRE	Interferon stimulated response element
ITAM	Immunoreceptor tyrosine-based activation motif
JAK	Janus kinase
KEGG	Kyoto Encyclopedia of Genes and Genomes
KNN	K-Nearest-Neighbors
LOWESS	Locally Weighted Super Smoother
LPS	Lipopolysaccharide
MAD	Median Absolute Deviation
MAR	Missing At Random
MCAR	Missing Completely At Random
MCMV	Murine CytoMegaloVirus
MIAME	Minimum information About Microarray Experiments
miRNA	micro RNA or microRNA

MOI	Multiplicity of Infection
MPS	Mononuclear Phagocyte System
NCBI	National Center for Biotechnology Information
NMAR	Not Missing At Random
NRMSE	Normalised Root Mean Squared Error
PCA	Principal Components Analysis
PCIT	Partial Correlation and Information Theory
POE	Probability Of Expression
REM	Random Effects Model
RMA	Robust-Multiarray-Average
ROC	Receiver-Operating Characteristic
RP	Rank Product
SCAP	SREBP cleavage activating protein
SREBP	Sterol Regulatory Element-Binding Protein
STAT	Signal Transducer and Activator or Transcription
TFBS	Transcription Factor Binding Site
TLR	Toll-like Receptor
VSN	Variance-Stabilising-Normalisation

Chapter 1

Introduction

This chapter provides background for understanding the use of statistical meta-analysis for the detection of subtle gene transcription changes in macrophage type II interferon signalling. This covers macrophages, interferon gamma signalling in macrophages, microarray technology and meta-analysis. Also included is a final overview of the thesis rationale, research context and structure.

1.1 Macrophage biology and interferon signalling

Macrophages (MØ) are an important and highly heterogeneous component of the innate immune system, where their primary function is the first-instance recognition, ingestion and destruction of pathogens or tumour cells. Their immune functions include signalling and antigen-presentation to other immune cells (e.g. neutrophils, lymphocytes), therefore also giving them a function in the adaptive immune system. In line with their first-responder immune function, macrophages can be found throughout many tissues, their wide distribution aiding recognition of any foreign material. Apart from immune functions, macrophages have also site-specific functions in wound repair and tissue organisation, steroid production, debris collection and homeostasis.

Background. The role of macrophages in absorbing and digesting pathogens by phagocytosis was first comprehensively (for the time) described in 1905 (Metchnikoff 1905). Subsequent research revealed much deeper levels of complexity in terms of function, location and development, leading to attempts at classifying macrophages within systems of cell types and subtypes (Aschoff 1924). These were superseded by consensus at a 1969 conference and with some revisions published in 1972 (Vanfurth, Spector et al. 1972). The proposed "mononuclear phagocyte system" (MPS) is close to modern understanding and

includes the system from origin (bone marrow precursors) over circulation (peripheral blood monocytes) to local application (tissue-specific macrophages). The definition of this system is based on a shared origin (bone marrow), function (efficient internalisation of large particles), features (cell surface receptors for immunoglobulins and pathogens), distribution (as monocytes via peripheral blood). The authors acknowledged that this concise understanding of the MPS would undergo continuous revision in light of new research.

In the decades since then, an overwhelming amount of knowledge has been generated (~288,000 macrophage-specific publications since 1972) on the development, distribution, morphology and function of macrophages. Phagocytosis in particular has been recognised as highly complex due to the wide range of different surface and internal receptors present in macrophages, as well as pathogen interactions with the macrophage itself (Aderem and Underhill 1999). This diversity of pathogen-recognition and -interaction routes accounts for complicated and interlinked signal transduction pathways. One of the most significant distinctions since made between macrophage subpopulations is that between classical activation and alternative activation, where either type has separate activating receptors (IFN-γ for classical; IL-4 and IL-13 for alternative) resulting in specific macrophage phenotypes with associated functionality. Classically activated and alternatively activated macrophages are also known as M1 and M2 macrophages, and the process of macrophages falling into either category as polarisation. In either case, these activation or pathways are usually balanced depending on type and time frame of infection, with a final macrophage deactivation stage applicable to both and induced by IL-10, TGF-β, glucocorticoids and other active molecules.

Classical macrophage activation. In case of bacterial infections, it was recognised early (Mackaness 1964, Mackaness 1969, Tripathy and Mackaness 1969) that the macrophages response to both reinfection by the same bacteria and subsequent infection by other bacteria was more effective than a passive macrophage reaction (innate macrophage activation by pathogens only) to the presence of bacteria would suggest, and that there were other factors (lymphocytes such as T-cells)

rendering a macrophage hypersensitive to bacteria, leading to fast and specific macrophage responses. While these investigations revealed that macrophages were somehow activated not only by the pathogen but also by host-specific factors, the identity of this factor as IFN-γ was determined much later (Nathan, Murray et al. 1983, Nathan 1991). This introduced the understanding of "classically activated" macrophages, where macrophages are primed (but not activated) by IFN-γ and activated by the presence of a pathogen (indicated by Lipopolysaccharides, LPS). In this model, T helper 1 cells (Th1) and natural killer cells (NK) produce the cytokine IFN-γ, which binds to macrophage IFN-γ receptors and keys them to enter a primed state that is associated with increased aptitude for antigen presentation, production of pro-inflammatory cytokines, chemokines and phagocytosis in reaction to LPS-sensing. Heightened macrophage immune functions are then triggered when bacteria are recognised through binding of LPS in their outer membrane to a corresponding receptor (TLR) on macrophages. IFN-γ signal transduction is presented in more detail in chapter 5.

Alternative macrophage activation. If unchecked, classical activation of macrophages would lead to runaway immune responses to infection. This motivated the search for complementary signalling processes that might inhibit this excess. IL-4 (produced by Th2 cells rather than the IFN-γ producing Th1 cells) was initially identified as macrophage-activating in 1987 (Crawford, Finbloom et al. 1987) and its function described in 1992 (Stein, Keshav et al. 1992), where it was found to reduce macrophage production of pro-inflammatory cytokines, and to initiate a different activation state that results in expression of MHC class II antigens. As opposed to classical activation, no secondary (pathogen-provided) signal is required for macrophages to enter this state. In this model, Th2 cells produce IL-4 (and IL-13), which binds to MMR (macrophage mannose receptors) and initiates a different macrophage phenotype compared to that of classical activation, in this case resulting in reduction of pro-inflammatory cytokines, enhanced response to parasitic pathogens, and tissue repair.

Macrophage phenotypes based on classical and alternative activation were summarised in 2003 (Gordon 2003), with the classical activation resulting in

enhanced macrophage sensitivity, pro-inflammatory responses, tissue damage, cell-mediated immunity and general microbicidal activity; and alternative activation resulting in anti-inflammatory responses, allergic responses, tissue repair, anti-parasite responses and macromolecule-mediated humoral immunity.

Macrophage response to viral infection. Apart from their role in bacterial infection, interferons are also known for inducing and modulating an antiviral state, that is, the attack of the virus itself and induction of apoptosis in virus-infected cells. This includes interferon alpha and beta (IFN-α / β), referred to as type I interferon (Bach, Aguet et al. 1997, Biron 1999, Taniguchi and Takaoka 2002), as well as interferon gamma (IFN-γ), referred to as type II interferon (Muller, Steinhoff et al. 1994). Type I and type II interferon are described in a subsequent section. It has been noted that in this area the notion of macrophage polarisation into M1 or M2 type may be too restrictive. It is subject to large amounts of signal transduction crosstalk (Sica and Mantovani 2012), as various viruses are able to co-opt macrophage activity to their own ends. This increases the importance of identifying which factors can shift the balance between viral infection and host immune response.

The above distinctions on activation and interferon signalling have remained intact in intervening years, but great detail has been added in terms of signal transduction pathways available through the application of microarray and other high-throughput experiment platforms. Differences in whole-genome transcriptional profiles have been identified for M1 and M2 macrophages (Martinez, Gordon et al. 2006).

Interferons. Of particular relevance to this thesis are transcriptional pathways related to interferon signalling in macrophages, and here microarray and other platforms have enabled the creation of databases specific to cell responses to type I, II or III interferon signalling (Samarajiwa, Forster et al. 2009), identifying around 2000 interferon stimulated genes.

The relevance of interferons to viral replication was noted in 1957 (Isaacs and Lindenmann 1957), when an unknown factor (later identified as type I interferon) released from membrane tissue infected with inactivated influenza virus subsequently interfered with live influenza infections. Since then, three families of interferon protein have been identified (type I, II, III), where type I comprises IFN-α, IFN-β, IFN-ε, IFN-κ, IFN-ω; type II comprises IFN-γ; type III comprises IFN-λ, IL-28 and IL-29. The role of interferons in the immune system is primarily seen as antiviral (Samuel 2001), with virus-infected cells producing type I interferons and antigen-stimulated cells producing type II interferons. Released interferon proteins engage with their respective receptors (IFNAR for type I, IFNGR for type II) in other host cells to start the signal transduction cascade, which results in the host's antiviral response by directing the transcription of genes and translation of proteins (e.g. Zinc Finger antiviral protein (Karki, Li et al. 2012)) that interfere with viral replication and attempts by viruses to subvert this response as well as modulating other immune responses.

Macrophages possess distinct receptors for type I and II interferon, but the presence of type III interferon receptors has not been conclusively shown for mouse macrophages as yet, although it has been shown for human macrophages (Hou, Wang et al. 2009) and there is circumstantial evidence for macrophages responding to virus-induced IFN-λ (Melchjorsen, Siren et al. 2006). As described before, type II signalling is required for classical macrophage activation. Type I signalling is not a required signal for macrophage activation, but plays a prominent role in the induction of interferon stimulated genes and is required for the production of nitric oxide after microbe-triggered sensing of LPS (Vadiveloo, Vairo et al. 2000).

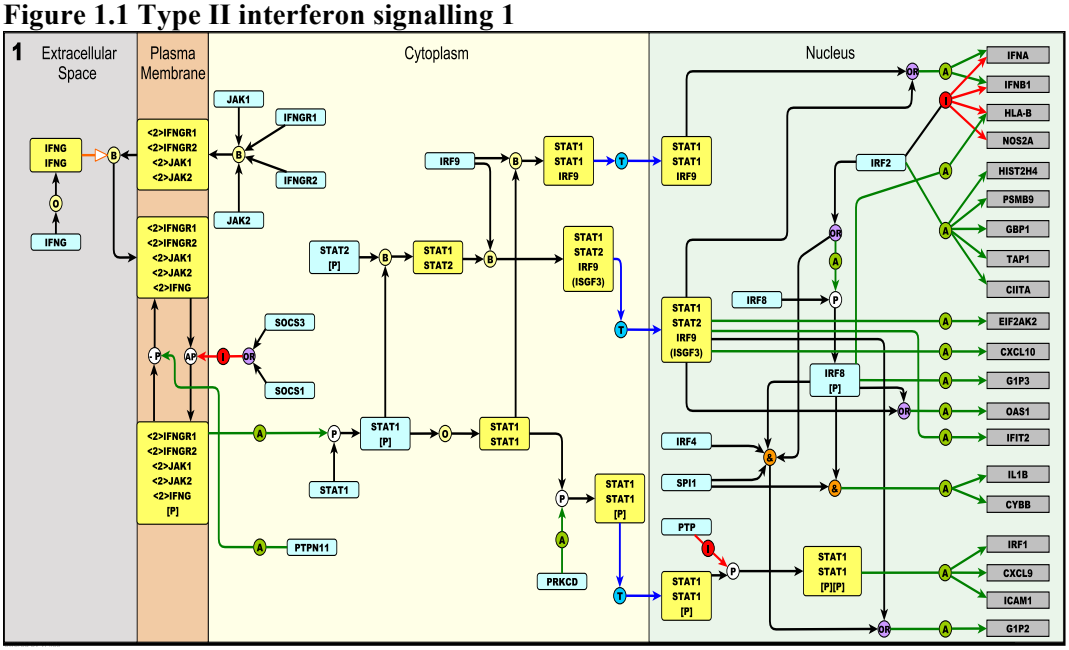
The three interferon classes described above have distinctive signal induction pathways (although crosstalk between them can be substantive) and are described below.

Type I interferon signalling is characterised by IFN-α or IFN-β (or all with the exception of IFN-γ) proteins binding to the interferon alpha receptor (IFNAR) on the cell membrane and activating the JAK-STAT pathway, resulting in STAT1/STAT2 heterodimers that can form a complex with IRF9 (then a

heterodimer), translocate to the nucleus and activate genes containing ISRE (Interferon Stimulated Response Element) motifs in their promoter. Type II interferon signalling is typified by IFN-γ binding to an IFN-γ receptor and also activating the JAK-STAT pathway, but where this results in the formation of STAT1 homodimers that translocate to the nucleus and activate genes containing GAS (Gamma Activated Sites) motifs in their promoter. Type III interferon signalling is less well characterised, and is assumed to share its function with type I interferon, but with JAK-STAT activation through IL28 or IL29 (and other cytokines) binding to an IFN-λ receptor. Both type I and type II interferon play a role in the activation of macrophages, although type II is regarded as the classical factor in this.

For the purposes of this thesis, it is important to note that the included microarray studies are all concerned with the effect of IFN-γ on murine macrophages in the absence of any other factors (infections or treatments). The expected primary response is therefore type II interferon signalling in macrophages in the absence of signalling through other immune cells, although the length of treatment will allow for autocrine and paracrine effects within the macrophage cell population. The classical activation of macrophages requires stimulation by an extracellular IFN-γ signal, which engages a cell's JAK-STAT pathway with the specific outcome of phosphorylating STAT1 proteins, which in turn dimerise, translocate to the nucleus and activate the transcription of genes containing a GAS (STAT1 homodimer binding site) motif in their promoter region. With no particular focus on the type of interferon involved, these genes and those affected indirectly are referred to as Interferon Stimulated Genes (ISGs). The type II interferon pathway has been previously developed into a logic map representation (Raza, Robertson et al. 2008) within the Division of Pathway Medicine (figure 1.1).

Figure 1.1 Type II interferon signalling 1



The pathway is arranged to flow from left to right. Components are coloured according to type (protein, complex or gene) and arranged within the sub-cellular compartments in which they are active. This pathway is initiated by IFNG binding to its receptor and a subsequent phosphorylation cascade involving a number of the JAK and STAT family of proteins. Several transcriptionally active complexes are formed (STAT1 homodimer, ISGF3 complex, STAT1:STAT1:IRF9 complex) and the pathway culminates with the transcriptional activation of target genes.
(*Figure 1 of Raza et al. BMC Systems Biology 2008, doi:10.1186/1752-0509-2-36, obtained under Creative Commons Attribution Licence*)

This canonical model does not include all possible modifiers of the type II immune response, and some of these are listed here.

- Other genes or miRNAs positively or negatively regulating (via respective proteins) the JAK-STAT pathway, such as SOCS, PIAS (Greenhalgh and Hilton 2001) or miR-155 (Lu, Thai et al. 2009) targeting and potentially down regulating SOCS (at least in T-cells), or nitric oxide synthase 2 (NOS2) associating with IFNGR1 in bacterial infection (Velez, Hulme et al. 2009). Other examples are *Prn*, *Ptpn1*, *Cd45*.
- Complicated crosstalk with other signalling pathways like MAPK signalling or PI3K-AKT (Rane and Reddy 2000), particularly type I immune response and in newer research, with *Ch25h* in sterol metabolism (Blanc 2013).

- Activation/suppression of gene transcription that is not due to STAT1 alone (such as activation of *Irf1* only through joint activity of STAT1 and TNF**α**-activated *NFκb* (Ramana, Gil et al. 2002). Other examples are *Jun*, *Cebpb*, *Hmga1*, *Myc*.
- Presence of a gamma-activated site in a gene may need to be complemented by both co-activators with STAT1, e.g. *Brca1* (Ouchi, Lee et al. 2000) and other transcription factors like SP1 bound to other sites on that gene (Ramana, Gil et al. 2002).
- In addition to interferon (or IFN-γ) related genes that are directly transcribed by STAT1, there are of course also genes (indeed, the majority of the interferon response) activated or suppressed further downstream and therefore only indirectly regulated by STAT1.

1.2 Microarrays

Microarrays are a physical substrate onto which nucleotide sequences are deposited as representation of genes, exons, or any other defined nucleotide sequence. When a biological sample is hybridised (binding of complementary nucleotide sequences) to a microarray, this allows the quantification of all genes' transcription in that sample.

1.2.1 Historical development

"Southern blots" were developed by Ed Southern in 1975 (Southern 1975) as a method to detect specific DNA sequences in a given DNA target (DNA extracted from cells or tissues). These can be regarded as the earliest precursor of microarray technology, although the term "microarray" was not officially used in publications until 1995 (Schena, Shalon et al. 1995) in a gene expression context. By this time, gradual advances had led to the use of RNA instead of DNA, substrate-fixing of probes rather than the target, larger numbers of gene probes, and substrate improvements. Although larger cDNA libraries had been screened before using filter paper and labour intensive processes (Kulesh, Clive et al. 1987),

the concept of large-scale deposition of oligonucleotide sequences on a physical substrate was first patented[1] in a series of patents by Ed Southern in 1998, although this has been subject to legal challenges (Rouse and Hardiman 2003). Various methodology papers followed (Maskos and Southern 1992, Maskos and Southern 1993, Southern, Casegreen et al. 1994), but Schena et al provide a first comprehensive application of microarrays in their use of glass slide substrates and high-density arraying robotics for dual-colour hybridisations, in this case 45 probes were robotically deposited in a ~20mm^2 area. While such glass-based oligonucleotide arrays were preceded by the earlier notion of photolithography (nucleotide sequences assembled base-by-base using masking technology) based sequence deposition (Fodor, Read et al. 1991), this technology did not become widespread until a larger number of gene sequences were known and could be included on more advanced versions, then known as Affymetrix[2] microarrays. Following the availability of fully sequenced genomes, 1997 saw the first whole-genome microarray study, an investigation of the yeast genome under multiple experiment conditions (Lashkari, McCusker et al. 1997). Since then, microarrays have become a mainstay of biological science, reflected in the number of newly published academic publications[3] rising from 55 in the one year period 1997-1998 to 5041 in the one year period 2011-2012. Although next-generation sequencing platforms (Metzker 2010) have since been developed into the next candidate for the collection of post-genomic data, the number of microarray data sets in public repositories (~34000 experiments in ArrayExpress[4] as of April 2013) provides a large amount of raw data open for re-analysis or collective analysis. Microarray technology is also unlikely to disappear soon or completely until next-generation sequencing can achieve sufficient levels of standardisation in processing, analysis and data handling (Biesecker, Burke et al. 2012).

---

1 United States Patent Office numbers: US6054270, US5700637 (note that these patents are later

2 www.affymetrix.com

3 Numbers obtained from Thomson Reuters Web Of Knowledge, searching for "microarray" in article title and "genes" in article topic (in order to exclude previous use of microarrays in electronic applications) and limiting period of search to the stated years.

4 http://www.ebi.ac.uk/arrayexpress/

1.2.2 Next-generation sequencing technology platforms

Next generation sequencing (NGS) technology has been commercially available since 2006 and has found increasing use in obtaining complete snapshots of genomes or transcriptomes at single nucleotide resolution. The approach itself has many technological implementations (Mardis 2008), e.g. Roche/454[5] sequencing, Illumina/Solexa[6] sequencing or SOLiD™ sequencing[7]. These differ in their sample processing protocols, labelling steps, data acquisition and performance; each platform also allows some degree of customising experimental protocols to highlight specific biological aspects (e.g. transcriptome sequencing). Generally, each method generates tens of thousands to billions of short reads. These are short sequences of nucleotide bases, the length and number of which is platform-specific. The total number of all short reads represents the full DNA or RNA sequence of a given biological sample, albeit unordered. In order to interpret the complete genomic DNA snapshot of the sample, these short reads are either aligned to a known reference genome ("resequencing") or aligned without a reference ("de novo sequencing"). Once aligned, the result is in theory the complete nucleotide sequence of all chromosomes in the original sample, although there are platform-specific limits and errors in the coverage and resolution of genomes. Platforms like RNA-seq[8] focus on the complete transcriptome by counting transcript abundance (frequency) of a given sequence in the sample. This provides similar outputs to microarray gene expression platforms, but in addition to providing expression level readings for known genes, it also allows the identification of novel genes or isoforms, splicing or single nucleotide polymorphisms. For analysis of RNA-seq transcriptome data, several challenges present themselves. The method and parameters of aligning short reads to a reference genome (as well as the choice of reference genome itself) affects the actual quantity measured, for example, one might measure splice variants of a gene independently or combined. Steps prior to statistical analysis therefore need

[5] www.454.com
[6] http://www.illumina.com/technology/sequencing_technology.ilmn
[7] http://www.lifetechnologies.com/uk/en/home/life-science/sequencing/next-generation-sequencing/solid-next-generation-sequencing.html
[8] http://www.illumina.com/technology/mrna_seq.ilmn

appropriate consideration. Regarding numerical outcome measurements, it has been shown (Mortazavi, Williams et al. 2008) that both microarray and RNA-seq transcriptome analyses match true sample transcript levels. However, while microarrays give continuous intensity readings (usually within a 16-bit range) based on fluorescence levels, RNA-seq simply counts short reads mapped to a genome. The latter can be treated as discrete data or, after some normalisation steps, as continuous data. Preferred analysis methods for RNA-seq have not yet been settled, and discrete (Poisson, Negative Binomial distribution), continuous parametric (Normal, Log-Normal distribution) statistical models as well as non-parametric models have all been used to similar effect in identifying gene differential expression (Anders and Huber 2010, Robinson, McCarthy et al. 2010, Busby, Stewart et al. 2013, Li and Tibshirani 2013, Soneson and Delorenzi 2013). While problems such as different signal ranges and sensitivity remain (Mortazavi, Williams et al. 2008), it is possible to combine or associate RNA-seq data with microarray data, and this has already been attempted (Battke and Nieselt 2011, Chavan, Bauer et al. 2013, Mooney, Bond et al. 2013). Very recently, meta-analysis methods were applied to multiple RNA-seq data sets (Jaffrezic 2013, Nikaido 2013), but the question of RNA-seq suitability for meta-analysis remains open, and combining RNA-seq data and microarray data within a meta-analysis may be possible but has not yet been explored. At this time, within the framework of this thesis any such exploration could only be theoretical, as no RNA-seq studies with the required experimental conditions currently exists, with the possible exception of a (albeit human) macrophage based study (Beyer, Mallmann et al. 2012). The meta-analyses in this thesis are therefore limited to microarray gene expression studies.

1.2.3 Microarray technology overview

Microarrays of the form relevant to this thesis (that is, discounting genotyping, exon, protein and other array types) consist of nucleotide sequences of assumed sufficient length to detect a complementary strand in the hybridised sample and to

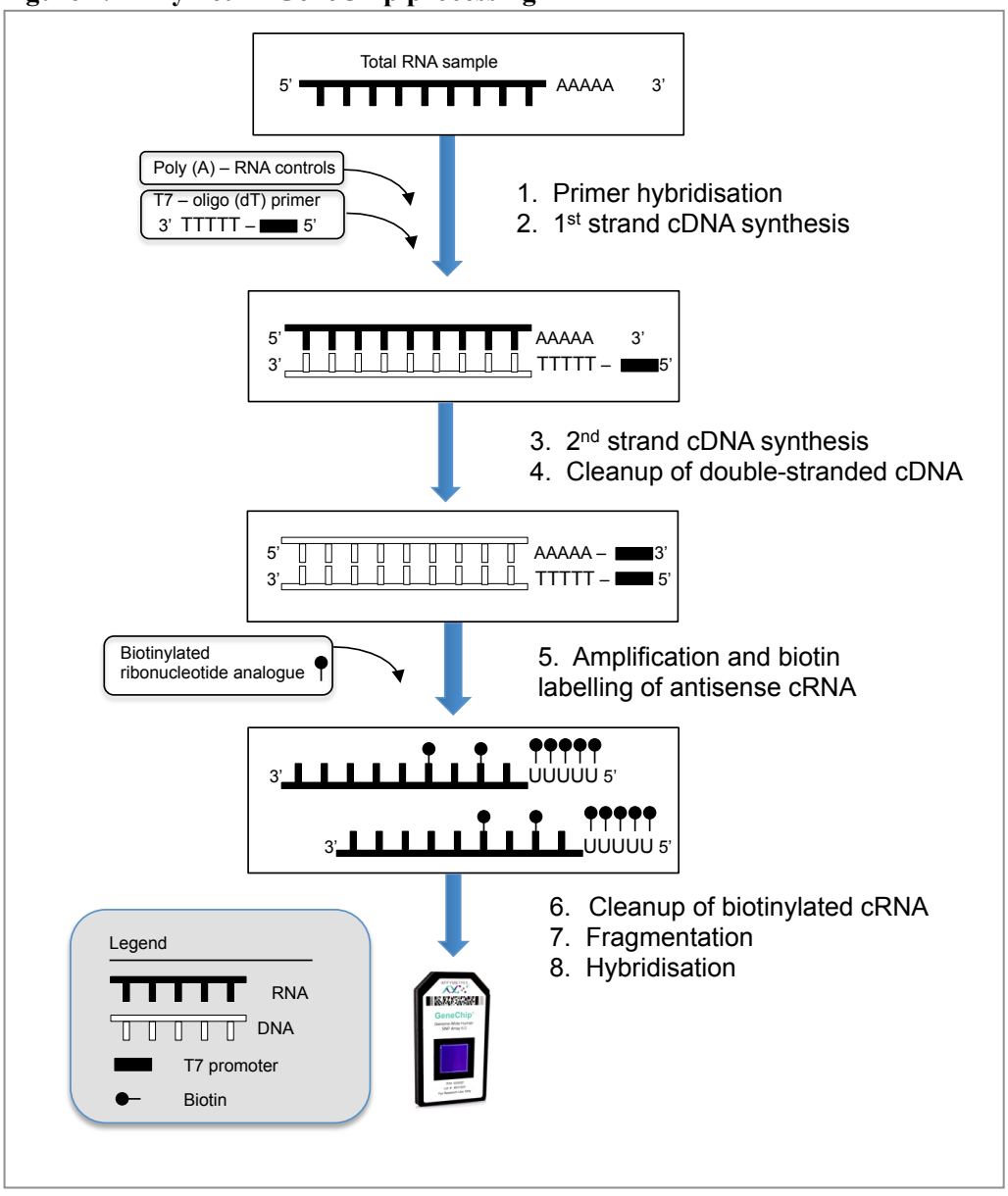
be representative of a given gene. The sequence probes located on the microarray are referred to as probes, gene probes or features. The biological sample hybridised to a microarray is referred to as target sample. All technologies have in common that in order to detect a gene that is transcribed in the target sample, it has to be labelled (usually with a fluorescent tag) in the target sample prior to hybridisation to a microarray, and only binding of probe sequence and complementary labelled strand in the target cause this label to remain present on the microarray. This fluorescence is detectable by exposure to the correct light excitation frequency in a scanner instrument, providing for each gene a measurement of its transcription ("expression") level in the target sample.

Among the basic differences between microarray technologies are the method of probe deposition on a substrate and the required labelling of sequences in the target sample. In "spotted" or "printed" arrays, probe sequences (multiple copies thereof) are directly deposited onto a predetermined location on the substrate with a robotic array printer. In an Affymetrix array or "GeneChip" (figure 1.2), individual probe sequences are assembled nucleotide-by-nucleotide in-situ with a photolithographic masking process, with probe copies distributed across the array rather than at a single coordinate. A third and more recent option are bead arrays (Illumina[9]), where probe sequences (and copies thereof) are bound to microscopic silica beads that then get randomly arranged in microscopic wells etched into silica slides. Microarrays can also be distinguished by the number of channels, where this is the laser frequency used to quantitate hybridisation levels for a particular dye label. Array platforms are usually either single-channel (single-colour) or dual-channel (dual-colour). The ability to read out this number of channels often but not always corresponds directly to the number of samples (1 and 2, respectively) hybridised to an array. Affymetrix arrays are single-colour arrays in that only one channel (laser frequency) can be quantitated from them. This usually means one biological target sample at a time is hybridised to the full array of probes, resulting in a quantitative absolute measurement of expression for each gene. Printed arrays are commonly but not exclusively used in a dual-colour setup, with a test and a control target sample (differently dye labelled) co-hybridised to

[9] www.illumina.com

one probe array and two channels quantitated. This results in a quantitative relative measurement of expression for each gene. Illumina bead arrays can be quantitated as single-colour or dual-colour platforms, depending on the particular type.

Figure 1.2 Affymetrix GeneChip processing

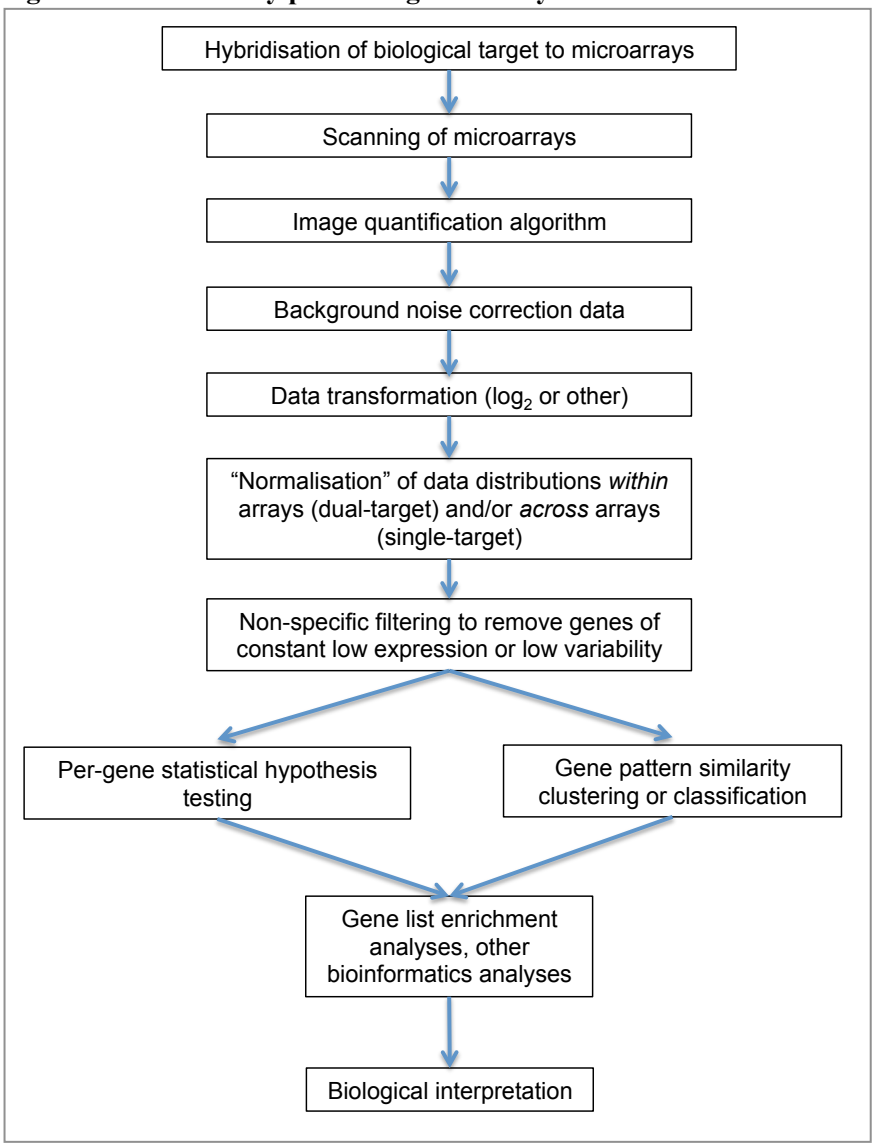


Depicted are all laboratory procedures that are prior to confocal laser scanning of hybridised microarrays and subsequent to growing bone marrow derived macrophages in culture and either treatment or not with IFN-γ.

1.2.4 Microarray processing and analysis

Subsequent to hybridisation of biological target samples to a microarray, a distinct number of steps are carried out before biological results are obtained. These are outlined in figure 1.3.

Figure 1.3 Microarray processing and analysis



This general workflow and methods applied at each step were introduced in the first few years after the introduction of microarray technology, with computational, statistical, bioinformatical and biological researchers applying their

existing expertise and specialisation to this new field. Although there are large numbers of later refinements and additions, the steps in this figure comprise the groundwork they are built on. A workflow based on sequential application of techniques is discussed in a paper on standard operating procedures for microarray analyses (Forster, Roy et al. 2003). Statistical issues regarding quantification of scanned microarray images are discussed in Glasbey et al (Glasbey, Forster et al. 2007). Issues requiring early solutions were centred on several subject areas, which are outlined below. It should be noted that these were mostly concerned with printed arrays because the commercial availability and use of Affymetrix arrays dates later than the availability of the former. While procedures and algorithms for stages of a microarray study (data acquisition, pre-processing, experiment design and analysis) may differ for that platform, the stages themselves are the same.

Data acquisition and storage. Microarray data provided amounts of data that limited the use of spreadsheets or text files for data handling and storage, making it necessary to develop and standardise methods for obtaining, identifying and storing data. Recognition of these issues was a primary motivator for establishing the Microarray Gene Expression Data society (MGED). From these efforts of developing standards came an early paper (Brazma, Hingamp et al. 2001) outlining the MIAME-standard (Minimum Information About Microarray Experiments) for characterising microarray studies and data, with the express purpose of making microarray data and their analyses reproducible within this framework of defined processes and terminology. This standard is still one of the underpinnings of the data collection and storage by the European Bioinformatics Institute's ArrayExpress database and the National Centre for Biotechnology's Gene Expression Omnibus database, two main providers of publicly available microarray data sets. It also paved the way for researchers to upload experimental data to a central resource, allowing easy and unprecedented access to data to re-analyse them, or as is the case in this thesis, to combine them.

Data pre-processing. Numeric microarray gene expression data are based on the quantified (using image processing algorithms) fluorescence of images scanned at the excitation frequency of the label used for tagging sequences in the target sample. Before analysing these data, it was recognised that further processing had to be applied in order to address noise and variation introduced by prior steps (e.g. RNA extraction, RNA amplification, dye labelling, hybridisation, scanning, array imperfections). In this regard, "normalisation" is now the step that removes such systematic sources of variation based on assumptions. The goal is to centre data distributions on 0 (if using dual-colour arrays that result in log-ratio expression measurements) or to match overall array distributions between all arrays in the study (if using single-colour arrays). For printed arrays, an additional issue were expression differences introduced by robotic arraying, with different "print tip" groups possibly leading to different measurement levels. In either case, normalisation is based on one of two fundamental assumptions. One is that the effect of any sample treatment or measured condition is limited to a small set of genes and not the full genome represented on the array. In case this cannot be assumed, the second assumption is that a selected set of probes (spiked in or existing housekeeping genes) should not change expression in response to treatment or in different conditions.

Although some early approaches (Kerr and Churchill 2001) suggested statistical analysis models (specifically, linear models incorporating variation sources) could account for all types of variation and error, independent normalisation alternatives were proposed and later incorporated into standard workflows. Although primarily concerned with gene expression profile clustering algorithms, normalisation strategies to address sources of systematic variation were usefully reviewed by Quackenbush in 2001 (Quackenbush 2001) and further expanded in 2002 (Quackenbush 2002). These early normalisation approaches were focused on printed dual-colour arrays (with expression data obtained as ratios) and centering log-ratios for all genes on an array around 0. Although this methodology does not apply to single-colour microarray platforms like Affymetrix or Illumina, the underlying assumption of a majority of genes on an array being unaffected by any tested condition or applied treatment remains common. For dual-colour arrays,

non-linear normalisation methods (e.g. fitting a weighted smoother line in a regression of signal intensity to signal ratio) remain common due to dye-label bias, but more generic methods were developed specifically for single-colour platforms. Primary examples of those are Variance-Stabilising-Normalisation (VSN) introduced by Huber in 2002 (Huber, von Heydebreck et al. 2002) and the Robust-Multiarray-Average (RMA) introduced by Irizarry in 2003 (Bolstad, Irizarry et al. 2003, Irizarry, Bolstad et al. 2003, Irizarry, Hobbs et al. 2003). The former uses the verifiable assumption that the coefficient of variation at low expression levels is different from that at higher expression levels, motivating different data transformations along the range of expression values. The latter incorporates a background correction, log-transformation, robust multi-probe averaging and is in effect a quantile normalisation, that is, it matches all quantiles (as opposed to just the median or another single point estimate) across all arrays in the study.

Experiment design and statistical analysis. Microarray technology initially had uncertain properties and characteristics, requiring re-establishment of common statistical procedures for this platform. This particularly concerned dealing with sources of variation, which included arrays, probes, dye labelling, RNA processing, hybridisation, target samples and numerous other potential sources. These questions were initially addressed 2001 (Kerr and Churchill 2001) and 2002 (Churchill 2002, Yang and Speed 2002), discussing differences between technical and biological replication, arrangement of dye labelled samples amongst the full set of microarrays, utility of universal reference samples, and various statistical study designs already well-known outside the microarray community.

The issue of study design is closely related to that of data analysis, and early work in this area still informs statistical microarray analysis models in use today. Concerns here included quantitative ways to characterise signal noise for inclusion in linear models, assumptions about the normal or lognormal data distribution of gene expression measurements, non-linear effects of dye labelling on gene expression, and the use of reference samples in a linear model. These were formally addressed in 2000 (Kerr, Martin et al. 2000) and resulted in the application of analysis of variance (ANOVA, already a well-established analysis

model in other areas of science) to microarray data. The linear model for determining the difference between group means (arithmetic or geometric) --given the measured within-group and across-groups variance and error-- was the theoretical foundation for applying per-gene hypothesis tests on differential expression. An important issue yet to be addressed were issues of "multiple testing", which is inherent in microarray technology. It is statistically inevitable that the testing of multiple hypotheses at the same time will incur high false discovery rates (that is, false rejections of the null hypothesis in statistical inference test). For microarrays, genes were recognised to be the variables on which inference tests are performed, and their number exceeded that common to other applications (e.g. clinical trials, where maybe a handful of physiological variables might be tested). Dudoit et al recognised in 2002 (Dudoit, Yang et al. 2002) that multiple testing adjustments to computed p-values for genes were a necessity in order to avoid large numbers of false positive results to validate. Investigations showed that the simplest approach (Bonferroni correction) would be extremely conservative in the presence of thousands of variables and presented several alternatives, including the less conservative and now frequently used methods of controlling the false discovery rate (FDR). These FDR methods were subsequently investigated in detail and confirmed as a viable choice (Reiner, Yekutieli et al. 2003).

Also of interest in that period of time were suggested alternatives to simple linear models, specifically those based on Bayesian statistics. These were discussed by Efron as early as 2001 (Efron, Tibshirani et al. 2001) as well as Smyth in 2004 (Smyth 2004) and laid the foundation for the now widely used "empirical Bayes" hypothesis tests, which -- instead of using information of one gene at a time – borrow information (in this case expression variance estimates) from other genes, resulting in more robust results for microarray studies consisting of small numbers of samples per group. A relatively late addition was the Rank Product method (Breitling, Armengaud et al. 2004), a simple yet robust statistical test focusing not on group mean expression differences, but on relative rankings of differential gene expression that take into account broad relationships between genes in terms of

their rank consistency across all samples (or rather, sample combinations of treated and control samples).

In parallel to statistical hypothesis testing, the development of pattern recognition algorithms for gene expression profiles began in 1998 with Eisen's seminal paper on cluster analysis (Eisen, Spellman et al. 1998). This began still on-going attempts to assess similarity (or dissimilarity) patterns between genes in their expression levels changes across multiple samples, usually multiple biological conditions or sequential time points. Using machine learning algorithms to identify such clusters of gene expression is aimed at identifying gene function modules on the basis of gene co-expression. Such measurements of similarity or association between genes are also used for generation of gene network graphs, where genes are drawn as nodes and the strength or nature of their relationship is indicated by a drawn edge between them (Freeman, Goldovsky et al. 2007).

It was specifically the application of clustering algorithms that highlighted the issue of missing values in microarray data sets, since many algorithms require the presence of complete data matrices to calculate distances between genes or samples. This resulted in many algorithms designed to estimate replacement values (to "impute") based on the values of other genes or samples, primarily through first finding genes or samples similar to the one with missing values. These are in common use, but they have never been explored in context of hundreds or thousands of missing values that are introduced when combining different microarray platforms (with differing genome coverage) for a meta-analysis. Chapter 4 introduces and addresses the issue of missing values in detail.

Microarray technology presented a necessity and opportunity in the measurement of full or large proportions of genomes. It provided a wealth of information previously unobtainable in smaller-scale experiment platforms, but it also required difficult large-scale validation of all positive results. Into this space fell the development of bioinformatics tools and resources. These provided existing knowledge in form of gene annotation, e.g. identities, functions, pathways, or related research. Increasingly, they also provided analytical tools for testing gene

lists (usually resulting from prior statistical or machine learning analysis) for enrichment of particular gene functions, membership of known pathways or other properties. One important early resource providing existing annotation in form of a database was DAVID (Dennis, Sherman et al. 2003), tools like GOstat (Beissbarth and Speed 2004) added to the field the capacity to perform the above mentioned enrichment analysis of existing annotation. These tools have proven invaluable, undergone continuous development and now exist amongst a very large number of bioinformatics tools at the disposal of researchers to make sense of and create new hypotheses from initial analyses of microarray data.

1.2.5 Transcriptome

DNA microarray technology is used to measure a biological sample's transcriptome as a proxy for biological function carried out through the proteome. Since the conception of this technology, many discoveries have been made that highlight the its limitations. It is known that, firstly, mRNA (messenger RNA) expression levels in many cases do not have a proportional relationship with protein abundance or the magnitude of their biological effect (Ozbudak, Thattai et al. 2002, Watterson, Guerriero et al. 2013). Secondly, it is known that mRNA comprises only part of the transcriptome, and microarray technology by its design (with the exception of tiling arrays) will measure expression of protein-coding genes. However, there are a multitude of non-protein coding RNAs that have been found to interact with themselves or other RNAs and proteins (Wan, Kertesz et al. 2011). Examples of these non-coding RNAs are miRNA (microRNA), siRNA (short interfering RNA), rRNA (ribosomal RNA) and tRNA (transfer RNA), and their total number (in the human genome) has been estimated at more than 400,000 such RNA sequences (Willingham and Gingeras 2006), although it is unclear how many of them have an actual function (Rederstorff, Bernhart et al. 2010).

In the above context, RNA-seq as a full-genome coverage technology has been a useful tool (Forrest and Carninci 2009, Fasold, Langenberger et al. 2011) in identifying and characterising non-coding RNAs.

Within the framework of this thesis and for any meta-analysis on data obtained from databases, it is important to acknowledge that these data are often from older generations of microarrays and that this means the measured mRNA levels do not represent a complete picture of expression. The design of these arrays will miss both protein-coding (mRNAs not known or predicted at the time) and non-coding RNAs, and the gene sets identified through any analysis (including the meta-analyses in this thesis) will only represent a limited starting point to understanding the interactions of transcription, translation and post-translational events.

1.3 Meta-analysis

A detailed introduction to the specific meta-analysis models used in this thesis is provided in chapter 3, including historical context and equations. The introductory section here is therefore limited to providing the general context of meta-analysis and ultimately its relation to microarray studies.

1.3.1 Definition and background

A contemporary definition is provided by the Cochrane Collaboration[10] glossary, with meta-analysis as "The use of statistical techniques in a systematic review to integrate the results of included studies", where a systematic review is a "review of a clearly formulated question that uses systematic and explicit methods to identify, select, and critically appraise relevant research, and to collect and analyse data from the studies that are included in the review".

A more concise definition from a statistician's point of view (Dersimonian and Laird 1986) is "the statistical analysis of a collection of analytic results for the purpose of integrating the findings".

[10] www.cochrane.org (This is an international collaboration serving as a centralised resource for collecting officially registered randomised controlled clinical trials, meta-analyses, support and information)

In either definition, it is of particular importance that meta-analysis is not concerned with combining the observed *data* from multiple studies, but with combining the *results* obtained in multiple studies.

Background. The idea of combining several studies on a subject to obtain a better result is not new and has existed for centuries outside of any formalised framework, findings were anecdotally or selectively combined or limited to literature reviews. Although the latter are related and useful tools for combining the knowledge inherent in a field, they lack the quantitative and statistical framework to formally combine numerical findings across studies. The first documented use of a formal quantitative combination of results from multiple similar studies can be attributed to Karl Pearson, who combined typhoid inoculation studies on soldiers in order to assess correlation between typhoid disease, mortality, and inoculation (Simpson and Pearson 1904). The basic principle used in that report consists of computing estimates for the strength of association between two variables for each study, with an overall estimate then calculated across all similar studies. Pearson noted that the variation of per-study estimates had a large effect on the outcome and suggested formal investigation of such variation across studies. Although not named in the original report, this introduced the concept of study heterogeneity and is of primary importance in meta-analysis. This initial report has given rise to decades of research on this subject, and it inspired the concepts still in use for general meta-analyses today. The term “meta-analysis” itself was introduced in 1976 (Glass 1976) and it is one subtype of approaches collectively known as research synthesis (Chalmers, Hedges et al. 2002). “Meta-analysis” maintains its special definition, but if used by non-statisticians, “meta-analysis” is occasionally used as collective term for other forms of research synthesis, which include “data integration”, “data fusion”, and others that combine data rather than results or that are another form of data integration. This thesis is concerned with the statistical form of meta-analysis and can draw no conclusions about the effectiveness of other research synthesis approaches.

Objectives. The main objective of many meta-analyses is to obtain a quantitative *combined estimate* of the effect of a treatment or condition on a measured variable. For clinical trials, this may be the observed size of the effect of aspirin (the treatment) on cancer patients' mortality rate (the variable). It should be noted that in context of microarrays, each gene is a variable, which requires that as many meta-analyses are performed as there are genes.

The underlying assumption is that any single study will provide limited statistical power to obtain meaningful results, or that a projected treatment effect is very small, or that the variable of interest is heterogeneous between study populations. A combined estimate in theory provides larger virtual sample sizes (where sample size here refers to the number of biologically independent samples, i.e. individual mice) and therefore a more powerful statistical representation of the treatment or condition under investigation, although this is subject to many caveats regarding selection and design of suitable studies and analysis methods. A secondary objective is often an active investigation into how different studies contribute to overall estimates of treatment effect, that is, study heterogeneity itself is the subject of interest.

1.3.2 Meta-analysis key considerations

For a meta-analysis to be successful, correct selection of studies and application of analysis is crucial but subject to many biases (Egger and Smith 1998, Walker, Hernandez et al. 2008). Key considerations are here detailed, also taking into account how these relate to meta-analyses based on microarray technology.

Publication bias. Of foremost concern for meta-analysis is publication bias, in which the availability of studies is biased towards those with positive results, that is, studies with negative outcomes are often not published, severely limiting the interpretability of meta-analysis results. Interestingly, for microarray studies this problem is potentially alleviated for basic biological studies, because measuring full genomes can imply that at least some genes will observably respond to some intervention or show differences in some condition, increasing likelihood of

publication. Informal investigations (Ramasamy, Mondry et al. 2008) suggest that a small percentage (10%) of commercially purchased microarrays do not get published, although it is unclear if this is caused by negative results, technical array failures, or other reasons for non-completion of project.

Search parameters. Even if a study has been published (and in case of microarray studies, data uploaded), a successful meta-analysis requires that the underlying investigative question is similar enough across studies. Initial searches depend on the tools used (literature databases, search engines, study registries) and the search criteria chosen (keywords). For microarray studies, biological domain knowledge is of crucial importance, as biological experiments are often removed from a (relatively constrained) human subject setting and the exact nature of organism, tissue types, cell types, treatment regimes and technology platform become very important.

Study selection bias. Once search parameters are decided and candidate studies identified, a selection needs to be made on the basis of inclusion and exclusion criteria for studies to undergo meta-analysis. This concerns study design (randomised controlled trials, group sample sizes, the type of replicated samples (individual subjects, cell lines, pooled samples, technical replicates) source population, time frame and more), questions of data quality (inclusion of low quality studies will have an effect on combined meta-analysis outcomes) and processing and availability of data (are per study effect sizes of the same type and the same scale?). For microarray studies, study design is of the same importance, but with regard to data quality and data processing they present a sizeable advantage in the availability of all unprocessed data in public repositories. That is, for microarray meta-analyses one is not limited to effect size or p-value estimates listed in published articles, because the full unprocessed data sets can be downloaded and individually and identically processed and analysed.

Study heterogeneity. It can be assumed that no two independent studies will return absolutely identical results. For meta-analyses, this poses a problem in terms of the

degree of the dissimilarity between study results (heterogeneity). Considerations here are how similar primary study results of the observed treatment effect have to be for a study to be included in a meta-analysis. In the absence of knowledge on the cause of any differences (which may be real differences, or issues with population sampling, treatments, analysis), a decision usually needs to be made based on a metric. One such method is a forest plot (Lewis and Clarke 2001), a reference example of which is shown in chapter 3 (figure 3.2). This plot depicts the effect size and its confidence interval (CI95, meaning the true population effect size will lie in this range with a probability of 95%) for each individual study, allowing a rough decision to be made based on the amount of overlap between these confidence intervals. That is, for studies to be selected, their confidence intervals should ideally overlap and share the same effect directionality (i.e. either a positive or negative effect size). However, this can only form the basis to visually identify extreme outlier studies. An alternative to this visual assessment is a statistical test for study heterogeneity (e.g. a χ^2 test based on the null hypothesis of "no differences in effect size across all studies"), although this is also not free of risk: if there are too few studies, the statistic is flawed, if there are too many studies, the statistic is too good at identifying small heterogeneity levels. In terms of microarray studies, the issue of heterogeneity is difficult to address, because the number of studies and the size of studies are usually small (this is quantified in chapter 3).

Study size. Small study sizes mean that the statistical power of individual studies to detect small expression changes is very limited and subject to error, increasing heterogeneity between studies. The limitation in numbers of studies prevents reasonable estimation of study heterogeneity as discussed above. This is a clear limitation compared to meta-analyses on randomised clinical trials, where studies are usually required to be sufficiently powered and where more studies may be available. The use of forest plots or heterogeneity tests is also limited because in addition to plotting per-study outcomes, this would also have to be done for each gene (variable) separately. An investigation into summary statistics (e.g. proportion of genes on array for which confidence intervals do not overlap across

studies, or for which a test indicates significant heterogeneity across studies) would be of worth, but for this thesis the issue is considered secondary to biological validation of meta-analysis usefulness.

Multiple testing issue. Identical to the multiple testing issue described in section 1.2.4, microarray meta-analyses in particular are subject to a large number of statistically significant but false positive results because many thousands of variables are tested simultaneously. A 1-in-20 accepted false positive rate for a single test implies 500 false positive results when testing 10000 genes, and if required for publication reasons, meta-analysis p-value thresholds can be adjusted downwards as explained in section 1.2.4.

1.4 Microarray meta-analysis

A detailed introduction to meta-analysis models used in this thesis is provided in the introduction to chapter 3, including historical context and equations. This section introduces the field of microarray meta-analysis, because this has advanced independently of the general concept.

1.4.1 Literature summary

The field of microarray meta-analysis is small compared to many other areas. Using strict search criteria requiring the title of a published article to contain both "meta-analysis" and "microarray" identifies 116 journal articles in a Thomson Reuter Web of Knowledge search across all publication years (as of 19 Mar 2013, shown in figure 1.4) The resulting body of publication begins in 2002, around the same time as research into general microarray analysis issues (section 1.2 above). There are clear indications that, as a research theme, this subject peaked in 2008 and subsequently is associated with less primary research interest (although meta-analyses continue to be applied and user-oriented software continues to be developed). It is unclear if this is related to statistical researchers changing their focus to analysis issues around next-generation sequencing platforms, although a

corresponding search for "next generation sequencing analysis" (as individual search terms) provides an indication that this may be a contributing factor (figure 1.5), with this subject rising after microarray meta-analysis peaks. At the outset and during the development of a body of research for this thesis, the subject matter of microarray meta-analysis would have presented an ideal opportunity to contribute to the first wave of developments, it is therefore unfortunate that the extended part-time nature of this thesis prevented timely use of this opportunity. However, as shown in figure 1.4, this work will still fit into (albeit a reduced quantity of) contemporary research, and it is a positive that the actual use of microarray meta-analyses on biological data is still on the increase, assuming a search of the terms "microarray" and "meta-analysis" as a *topic* rather than as part of a publication title presents suitable proxy evidence for this (figure 1.6).

Figure 1.4 Publications with "microarray" and "meta-analysis" in title

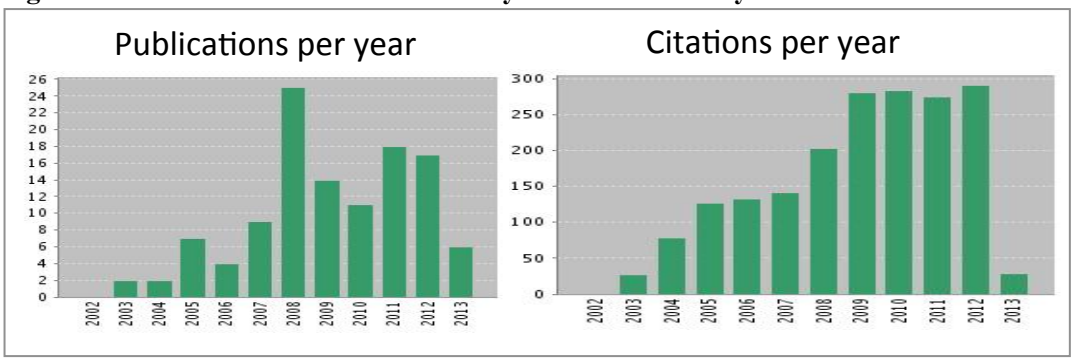


Figure 1.5 Publications with "next" "generation" "sequencing" "analysis" in title

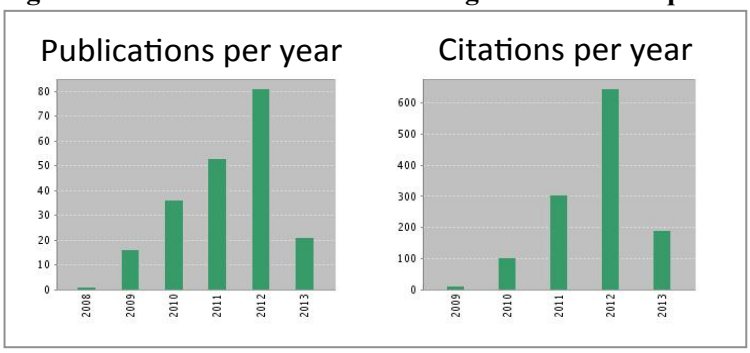
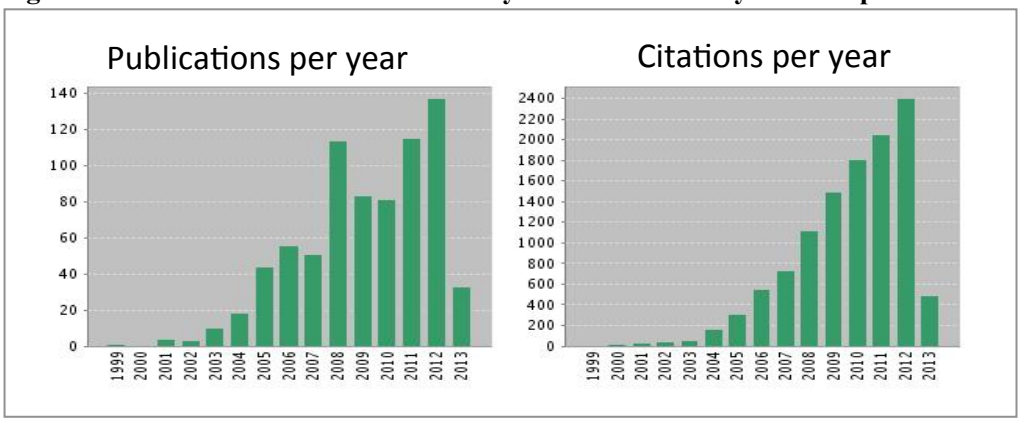


Figure 1.6 Publications with "microarray" and "meta-analysis" as topic



It is possible to speculate that the relative scarcity of microarray meta-analyses (for reference, there are >30000 published papers on meta-analysis of clinical or other *trials* within the last 5 years alone) is not only due to insufficient numbers of research studies, high diversity of research hypotheses, and reluctance to recreate similar existing data, but also to the absence of sufficient numbers of statistical analysts within biological departments, or the absence of easy-to-use software for meta-analysis purposes.

Despite the reduction of new content produced on meta-analysis methodology, this remains a current research subject (for microarray as well as next-generation sequencing technology) as well as an applied methodology. Current updates to the field are outlined in paragraph "Current Developments" in section 1.4.2.

1.4.2 Development of meta-analysis for microarray data

Earliest application to microarrays. The concept of applying statistical meta-analysis to studies based on microarray data coincides with the increased availability and affordability of microarrays (both printed and in-situ deposition arrays) and one of the earliest suggestions for this can be found in a 1999 paper (Khan, Saal et al. 1999). However, the first attempt to formally apply a meta-analysis by statistically combining the per-study significance test results was that performed by Rhodes et al in 2002 (Rhodes, Barrette et al. 2002). This study was

based on microarray studies aimed at characterising (differential) gene expression profiles between healthy prostate tissue and prostate cancer tissue derived from human clinical biopsies. Four such studies had been carried out independently, with between 8 and 23 samples available in the cancer groups (and usually fewer control samples). Although the meta-analysis model was unnamed, the model used is Fisher's sum of logs (Fisher 1932). This is one of the simplest and easily understandable methods (also used as a reference method in this thesis) in that it uses –for each gene— the per-study calculated p-values and constructs an overall meta-statistic by summing the log-transformed p-values. This statistic follows a chi-squared distribution and its statistical significance can be looked up in the relevant degrees-of-freedom table, although the authors here preferred to obtain significance by comparing to a permuted Null distribution based on the source data. The authors conclude that despite the presence of one outlier study, meta-analysis increased the significance and number of "called" genes (genes significant at chosen alpha level). Importantly, they were able to identify (although no formal count of such genes is provided) biologically relevant genes whose lower expression changes were not identified as significant in individual studies. Apart from proving that in principle meta-analysis can be of real benefit, this study already has the features that are common to microarray meta-analyses: full microarray data sets can be obtained and individual studies newly and consistently processed and analysed "in-house", a mapping of gene probes between arrays is necessary (here using UNIGENE IDs), and the resulting output requires new methods of visualisation and summary. As an early work in this area, some issues would require further discussion, such as the use of one-sided statistical tests in the individual studies, the use of UNIGENE IDs to ensure that gene probes on different microarray platforms represent the same gene, and of course the lack of reference meta-analysis methods. This study is therefore an early proof-of-concept by a biological research group, inviting more meta-analysis-centric research.

Systematic testing of methodology. This line was taken by Choi et al in 2003 (Choi, Yu et al. 2003), with the biological validation not the primary concern, but a systematic application of a more recent meta-analysis model (Hedges 1985)

based on a model-based combination of per-study effect sizes under fixed-effect-model (FEM) or random effects model (REM) assumptions on the heterogeneity of studies. This approach is detailed in chapter 3; the following simply puts it into context. This study used the same prostate data sets as those used by Rhodes et al above, and the methodology obtained a modest improvement over Fisher's method in identifying six additional and biologically relevant genes. This paper also proposed the introduction of an IDD (integration-driven discovery) metric and the related IDR (integration-driven discovery rate), highlighting that to assess the usefulness of a meta-analysis in a microarray context, one should compare how many discoveries (statistically significant genes) can exclusively be identified by meta-analysis and not by analysis of any individual study.

New methodology. Another relevant contribution to the field was a method described by Breitling et al in 2004 (Breitling, Armengaud et al. 2004), although this was not explicitly proposed as a meta-analysis model. Briefly (this method is detailed in chapter 3), the proposed Rank Product method is a non-parametric approach to microarray analysis. Instead of using parametric group mean comparisons (e.g. t-test) that utilise gene variation information, Rank Product uses the more directly biological metric of fold-change between individual case and control samples. By computing all possible fold-changes (which depends on the number of case and number of control samples in a study) for a given gene and ranking this with respect to all other genes, a summary statistic (the product of these rankings, which can be assigned a statistical significance) is obtained for each gene that indicates how consistently it is top-ranked across the replicate samples. While this applies to replicated samples *within* a microarray study, it is not difficult to see that this can easily be extended to consider replicated samples *across* multiple microarray studies, by simply forming the rank product across all per-study ranks of differential expression. The authors later explicitly described the methodology as capable of meta-analysis in a 2006 paper (Hong, Breitling et al. 2006) on the implementation of their algorithm as an R analysis package. As with any non-parametric analysis, benefits of this approach lie in its robustness against outlier observations (with ranks impervious to particularly high or low

values), which has perceived advantages when dealing with small numbers of observations or a high level of noise, both of which are issues with small microarray studies. This independence from the assumption of underlying data distributions (normal or otherwise) is balanced by a lack of statistical power when compared to parametric methods correctly applied to low-noise data. Of additional interest with the Rank Product method is that the relative expression levels between genes have an influence on the outcome, whereas parametric methods usually mean that genes are analysed independently of one another. Rank Product analysis was applied as a meta-analysis to a set of two microarray studies (Vert, Nemhauser et al. 2005), although little information is provided on its effectiveness.

Other "meta-analysis" methods. The microarray meta-analysis research described above informed the initial choices made in the hypothesis underlying this thesis. Separately, a wide variety of other types of research synthesis and meta-analysis continued to be advanced: Combining differential expression across studies (Ghosh 2003); combining expression profiles (Rhodes, Yu et al. 2004); correlating the within-study gene correlations across studies or "integrative correlation" Parmigiani's (Parmigiani, Garrett-Mayer et al. 2004); using a Bayesian model of combining fold-change estimates by treating other studies as the modelled prior knowledge (Wang, Coombes et al. 2004); effect sizes combined by a random effects model (Grutzmann, Boriss et al. 2005); combining binary "votes" across studies for genes based on a set of genes with known biology or "election results" (Benedict, Geisler et al. 2006); relying on researchers' published gene lists and simply assessing overlap of lists between studies (Cahan, Rovegno et al. 2007); cross-correlation and auto-correlation statistics between time-course studies of identical design (Keegan, Pradhan et al. 2007). These publications have in common that they are not performed from a formal comparative point of view. That is, it is not the meta-analysis approach that is compared to other meta-analysis approaches, but the meta-analysis proposed is simply applied to a data set of choice and compared to individual study outcomes. In the context of this thesis, the above studies were lacking a basic comparison of

existing meta-analysis models (from other research domains) to a proposed new approach, as well as the biological follow-through in terms of detailed quantification of meta-analysis benefits on biological discovery.

Comparative research into meta-analysis. This was the open research area into which this thesis was originally developed: given a specific biological problem (low-powered microarray studies on macrophage activation), can established meta-analyses make a quantifiable and verifiable biological contribution, how do methodologies compare and how can they be improved? These issues –with the exception of improvements to existing methodology-- were also recognised by Breitling et al and in 2008 they published a paper (Hong and Breitling 2008) addressing it, using the basic models described above (Fisher's sum of logs, effect size model, and their own Rank Product meta-analysis). There is therefore a definitive need to highlight the areas in which this thesis differs, these being a) notional improvements to the actual meta-analysis models, b) inclusion of more studies (other papers use effectively two small studies, with a third used only as reference), c) a specific biological system, d) result validation against biological domain knowledge rather than simulation or reference studies, and e) an investigation of missing values and assessment of methods dealing with them. On the subject of meta-analysis, this thesis is therefore aimed at a comprehensive assessment of the statistical models and potential improvements, but also the biological context of their results.

Current developments. While all literature discussed above describes the state of play in microarray meta-analysis that coincides with the main period of statistical work on this thesis, the field has not stalled and further developments have taken place. It is therefore important to assess how much of the relevance of this thesis can be retained in light of any more recent findings. The first comprehensive review of microarray meta-analysis has recently been published (Tseng, Ghosh et al. 2012), which considers 333 microarray meta-analysis papers, including those which do not develop methodology, but simply apply it to their biological area of interest. From these, they establish four classes of microarray meta-analysis

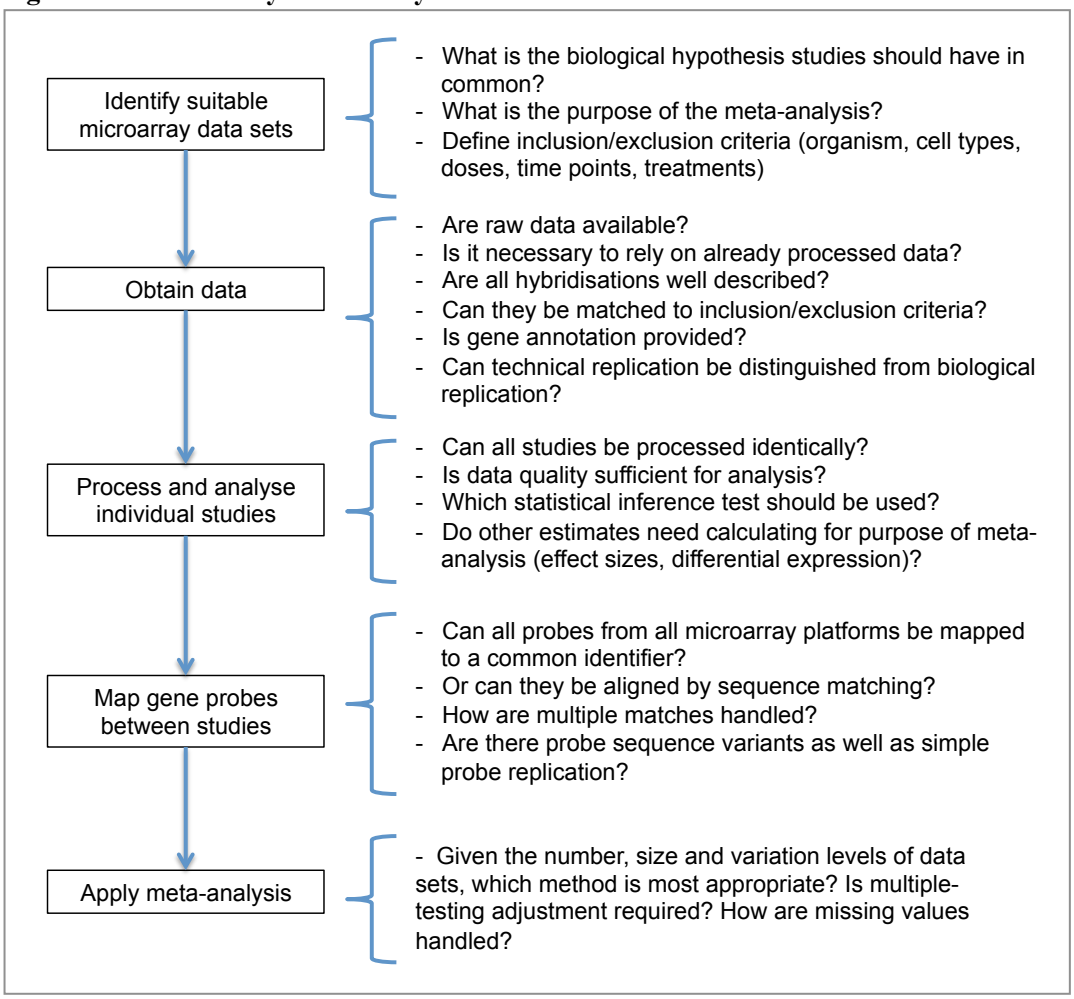
models: those based on combining p-values (42%), combining effect sizes (22%), combining ranks (9%) and those that merge data sets rather than combining results (27%). Three of these classes mirror those used in this thesis, with Fisher's sum of logs, effect size model and Rank Product model falling into the first three classes, respectively. The last method, merging of data (co-normalising all samples from all studies) makes a very strong assumption about the similarity of study hypothesis, experiment design and laboratory procedures across multiple studies. These methods use and integrate data directly, while statistical meta-analyses use indirect study outcomes (p-values, effect sizes, differential expression). While direct data integration is worthy of discussion, it has not been considered as a meta-analysis method for this thesis. Findings from this review suggest that there have been no major changes to the three methods used in this thesis (with the exception of a weighted form of Fisher's method) and that they remain in widespread use. The review also indicates that no "best" methods have as yet been established, due to lack of a large-scale and comprehensive comparative study. One interesting addition not covered in this thesis is a Bayesian approach (Choi, Shen et al. 2007, Scharpf, Tjelmeland et al. 2009) that does not use gene and study specific outcome representations (p-values, effect sizes, differential expression) that are then combined in a meta-analysis, but which uses a transformation (making use of all genes) of each microarrays' raw expression values to a "Probability Of Expression" scale (-1 to +1), allowing the meta-analysis to become a form of regular microarray analysis by virtue of making a new data set that consists of all POE-transformed microarrays from all studies. Overall, the current body of work in microarray meta-analyses is incomplete and there is scope to allow this thesis to contribute novel outcomes and model improvements.

1.4.3 Microarray meta-analysis workflow

Although inherent in the structure of this thesis, as an introduction it may be useful to summarise the steps required. While this is discussed in a large proportion of microarray meta-analysis papers, a specific and comprehensive summary is provided by Ramasamy et al (Ramasamy, Mondry et al. 2008). Figure 1.7 is a

summary graph to illustrate the normal processes, and it does not include steps required in an investigation of the meta-analysis itself. Complications and issues around these steps are presented in the corresponding chapters.

Figure 1.7 Microarray meta-analysis workflow



1.5 Thesis overview

Background and hypothesis. Gene transcription pathways related to macrophage activation have been extensively studied and gene transcription microarrays have been used by several independent research groups as a tool for investigating genome-wide macrophage-activation biology. This has resulted in a relatively

small number of studies based on similar biological hypotheses. A common characteristic of these individual microarray studies is their lack of statistical power to detect small expression changes (Yang and Speed 2002, Wei, Li et al. 2004), that is, the number of arrayed sample replicates is not large enough (given array sensitivity and financial or logistical constraints) to allow detection of genes with small transcription level changes. In addition to the statistical issue, this also often motivates use of arbitrarily set requirement for 2-fold or higher differential gene expression to be considered biologically relevant. It has been previously shown (through mathematical modelling) that small coordinated changes in expression level can affect cholesterol biosynthesis (Watterson, Guerriero et al. 2013), and that even genes with low transcription rates cause strong bursts of protein translation (Ozbudak, Thattai et al. 2002). The described experiment design limitations are pertinent to the hypothesis underlying this thesis, which posits that a) small changes in gene expression can be biologically meaningful and b) using statistical meta-analysis on the results of multiple similar microarray studies increases power to detect small but consistent gene expression changes induced by type II interferon signal transduction in macrophages.

A meta-analysis in theory results in genes for which differential expression between treated and control samples is too small to be statistically detected among the biological and technical variation inherent in any *individual* microarray study, but for which differential expression is *consistently* small across multiple studies. Although such genes may also be found through other experiment techniques (associated with higher sensitivity or easier processing of larger number of replicates), in terms of genome-wide assays the implication of missing meta-analysis approaches to this biological system is that numerous genes transcribed within or downstream of the type II interferon signalling cascade remain undiscovered.

On this background rests the hypothesis that responses of macrophages to immune activation by IFN-γ include small quantitative changes that are masked by noise but represent meaningful transcriptional systems in pathways against infection.

The existing body of work on type II interferon signalling is comprehensive and particularly work based on high-throughput biological experiment platforms has

already identified thousands of interferon-stimulated or potentially interferon-stimulated genes. Similarly, statistical meta-analyses of microarray data have shown some promise in contributing new candidate genes that in single microarray studies remain undiscovered because their subtle transcriptional changes are hidden within the biological variation and noise of individual microarray studies. However, to date no meta-analyses have been performed on type II signalling in macrophages, it is not clear which type of meta-analysis is best suited to this (or similarly limited) type of data, and it is unclear if any meta-analysis can provide meaningful biological results that add to the current knowledge on type II signalling. It is also an implicit purpose of this thesis to establish if acquired biological domain knowledge can alleviate the disconnect –present even in close collaborations-- between analyst and interpreter that disrupts the cycle of testing hypotheses and generating new hypotheses for further research.

Research objectives. The principal research question posed by the central hypothesis of this thesis involves four individual research objectives and questions. Firstly, can statistical meta-analysis provide meaningful biological results when applied to a heterogeneous (in terms of experiment design, not subject matter) and restricted set of small microarray studies? Secondly, can existing meta-analysis models be improved? Thirdly, can the imputation of missing gene expression values due to merging of different microarray platforms increase the scope of meta-analyses? Lastly, can the existing body of knowledge with regard to IFN-γ-mediated JAK-STAT signalling in macrophages be expanded based on meta-analysis results?

Methodology. In order to address the research questions outlined above, the following approach is taken. From an initial set of thirty identified microarray studies with relation to the biological system addressed in this thesis and found through local or public resources, six (only four of which are suitable for all meta-analysis models) microarray studies are selected based on key criteria, where these studies use mouse models, Affymetrix microarrays, and IFN-γ treatment of macrophages irrespective of dose or time point. Individual studies are analysed

before applying six different meta-analysis models to combine results across studies. Meta-analysis performance is first assessed quantitatively against existing knowledge. Possibilities for imputing (replacing, estimating) large numbers of missing values are investigated. Aggregated meta-analysis results are characterised in context of known macrophage immune signalling and new gene transcription hypotheses for the down-stream effects of IFN-γ-mediated JAK-STAT signalling developed and compared to an independent gene expression network analysis of a microarray time-course study.

Contributions to research. Within the applied methodology, this research makes a several novel contributions to the field. Meta-analysis benefits have previously not been investigated for IFN-γ-signalling in macrophages, or with a specific focus on their usefulness for limited collections of small studies outside standard data sets or simulated data sets. Improvements of the underlying meta-analysis models are here first proposed and tested. The influence of large-scale imputation of missing values is here tested for the first time. Existing knowledge of IFN-γ-mediated immune response genes is expanded through a high-confidence list of gene transcripts identified by meta-analysis, with speculation on their involvement in signalling crosstalk with type I signalling, sterol biosynthesis, and circadian rhythm. Finally, a gene association network analysis in substantiation of meta-analysis results shows the importance and biological implications of anti-correlated network structures.

Limitations. Every step in the workflow necessary to perform a microarray meta-analysis is associated with limitations that may affect the outcome of a meta-analysis in terms of quantitative results or interpretation. Meta-analysis is here tested within the framework of a small and well-defined biological question (although even this does not lead to homogeneous data sets), the selected studies are constrained to one microarray technology platform, and merging of genes across studies assumes that the same gene is represented in all studies although different probe sequences are used. The statistical analysis of individual studies and statistical meta-analysis of combined results does not exclude the possibility

that other model choices could provide a more powerful analysis, models used for this thesis were chosen in 2005/2006 and do not take full account of developments since then. The quantitative assessment of results as well as the biological interpretation make use of a limited number and possibly imperfect other reference sources, and interpretation bias is possible with regard to biological understanding of macrophage immune responses. It must also be noted that by definition, meta-analysis of multiple microarray studies can provide many true but not fully verifiable (without experimental characterisation of gene function) results, because it enables the detection of small gene expression changes that individual studies lack sensitivity for and which may therefore not previously be described in any relation to macrophages or immune responses to type II interferon activation.

Prior research. The effect of IFN-γ in macrophage activation has been repeatedly investigated through use of microarrays, with this research identifying around 30 studies on this precise or a closely related subject. However, this resource has not previously been used for meta-analysis, leaving a possibility that many genes with small but relevant transcriptional changes are as yet unidentified in this context. Previous attempts have been made to aggregate knowledge in this domain (such as the Interferome database) but this aggregation does not make use of formal statistical frameworks and may underutilise the power of combining study results). The three principal meta-analysis types used in this thesis were introduced as options for microarray meta-analyses in 2002, 2003 and 2006 (see section 1.4). These had a focus on different biological systems and did not involve model comparisons, although a comparison was later performed in 2008. Research into microarray meta-analyses has ostensibly peaked around 2008 but the numbers of applied microarray meta-analyses are still rising, indicating that interest in this subject will remain until next-generation sequencing (or other technology) has fully replaced the use of microarrays, but even in this scenario leaving large repositories of legacy data for future re-analysis or meta-analyses at no or small cost.

Thesis chapters. Subsequent to this introduction, chapter 2 introduces all data and methodologies used, with the exception of meta-analyses (covered in chapter 3). This includes the identification, retrieval and description of suitable microarray studies, the alignment of gene probes across multiple studies, processing and analysis of individual microarray studies, and biological and bioinformatical resources used. Chapter 3 introduces meta-analysis in context of microarrays, details the six meta-analysis models used, describes their application to the selected microarray studies and assesses the results quantitatively and in context of other data sources. Chapter 4 investigates missing value imputation strategies in an attempt to replace very large numbers of missing gene expression values that arise when merging microarray platforms with differing genome coverage. Chapter 5 aggregates representative microarray meta-analysis results and uses independent biological data resources and in-silico tools to characterise the results and speculate on new down-stream gene transcription hypotheses related to IFN-γ-activated macrophages. Chapter 6 applies a gene co-expression network analysis to an independent microarray time-course study that measures temporal gene expression changes of IFN-γ-treated macrophages, murine cytomegalovirus (mCMV) infected macrophages and combined treatment/infection. This is done in order to provide genome-wide expression patterns and modules of coordinated gene expression that can be related to the case-control type outcomes of statistical meta-analysis from chapter 5. Chapter 7 is a final discussion of overarching issues and suggestions for further work.

Chapter 2

Materials and Methods

All methods detailed here are listed chronologically in their order of application and cover the identification of suitable microarray studies, merging into a meta-analysis data set, initial data processing, statistical analysis of each study, and the generation of reference lists of known biology to compare meta-analysis results to. These methods do not include meta-analysis or evaluation of meta-analysis, these are described in detail in chapter 3 as they are the focus of the investigation for this thesis.

2.1 Microarray studies and data sets

2.1.1 Microarray data sources

For identification of microarray gene transcription studies relating to interferon treatment of biological samples, four resources were used:

1. ArrayExpress (http://www.ebi.ac.uk/arrayexpress/)
2. GEO (Gene Expression Omnibus, http://www.ncbi.nlm.nih.gov/geo/)
3. GPX MEA (Macrophage Expression Atlas, http://gpxmea.gti.ed.ac.uk/)
4. Journal manuscripts (PubMed, http://www.ncbi.nlm.nih.gov/pubmed/)

ArrayExpress and *GEO* are maintained as public data repositories by the European Bioinformatics Institute and the National Centre for Biotechnology Information, respectively. Both are MIAME (Brazma, Hingamp et al. 2001) compliant in terms of the type and amount of experiment design information gathered during the data acquisition process. Many journals publishing studies that include microarray data require or encourage the use of these two international resources to submit all expression data including detailed information on experiment design and laboratory protocols. These repositories were searched for key terms and the

results manually narrowed down to match the study parameters for this thesis, as described in section 2.1.2.

GPX MEA[11] is a microarray data repository with a specific focus on macrophages challenged with pro-inflammatory, anti-inflammatory, benign or pathogen insults. The data sets contained within this resource were compiled by staff of the Centre for Genomic Technology and Informatics (now Division of Pathway Medicine) at Edinburgh University and it contains microarray data sets generated by the centre itself or manually curated from academic collaborations or published articles that match the repository criteria.

Other data sources were identified through literature searches via PubMed, using all of the search terms (in turn, not simultaneously) "interferon" or "Ifn", "microarray", "macrophages" and the authors contacted directly with a request to obtain their microarray data.

2.1.2 Identification of relevant microarray studies

Studies were obtained from above sources in three stages. In stage one, generic searches for simple keywords were carried out on the sources (GPX-MEA excluded, as these were already known). A second stage removed redundant information and studies for which the search terms matched but were misleading. In stage three, studies with similar biological hypotheses and experiment design are identified as suitable for statistical meta-analysis.

Stage one. Resources described in 2.1.1 were searched for all microarray studies relating to search terms "interferon" or "ifn". Search terms for "macrophage" were not included in the database searches because of the possibility of missing studies where only specific macrophage cell lines are mentioned or where only shorthand forms (e.g. MØ) are used. The biological hypothesis underlying this thesis (that it is possible to use meta-analysis of multiple similar microarray studies to identify

[11] Disclaimer: the author of this thesis contributed advice and work on data processing and statistical analysis and advice on data base design issues to this resource.

gene expression changes not otherwise noted) requires identification of studies fitting a small range of criteria, and any automatic imposition of strong or ambiguous filtering criteria would be detrimental to this purpose. The above searches therefore retrieve all studies making mention of the quoted search terms irrespective of the context it is mentioned in. The consequence of relaxing search criteria is a large number of false positive study hits and a small number of suitable studies, requiring subjective but informed decision-making in a subsequent step. All searches were carried out in May 2008 and will therefore miss any relevant studies that have become available after this time[12]. Searching for "interferon" on GEO resulted in 215 original user-submitted microarray data sets, while ArrayExpress identified 85 (with 38 of those not already found within GEO. When searching for "ifn", GEO identified 127, ArrayExpress identified 42 (with 23 of those not already found within GEO).

Stage two. The total of 403 studies identified in stage one include multiple references to the same study and the majority are unsuited for consideration in a meta-analysis within the proscribed subject (macrophage activation), because they are studies for which the key search terms occur but are not directly related to study design. The total also contains studies using cell types that are not macrophage or dendritic cells. After removing all of the above (fibroblast studies were not excluded at this stage as it is an important modelling system), and adding four (known to be relevant to IFN-γ treatment in macrophages) studies from GPX-MEA and one study obtained through contacting the authors of a relevant paper with microarray results not yet submitted to a public repository, this stage results in a compilation of 30 studies that measure the effect of an interferon treatment on biological samples, at this time still irrespective of which organism this was done for, the exact treatment used, the microarray platform used for hybridisations, or type of biological sample (macrophages, other dendritic cells, fibroblasts) used. A

[12] Available in supplementary material, folder "List and description of searches for microarray data sets", files "MicroarrayRetrieverSearch…"

full list of these studies can be found in the supplementary material[13], this includes study identifiers and tabulated study experiment factors.

Stage three. For meta-analysis on a set of biological studies, the study hypotheses need to be similar. In this case, the biological hypothesis underlying the purpose of meta-analysis relates to the effect of IFN-γ (irrespective of dose and time points) on murine bone marrow derived macrophages. This limited hypothesis is also subject to assumptions that gene expression responses to stimulus would likely be different for different organisms, treatments or sample types. Therefore, this stage employs study inclusion and exclusion criteria in line with that hypothesis.

Inclusion criteria:

1. Study contains at least one group of samples that are bone marrow derived macrophages, treated with IFN-γ (any dose, any time point).

Bone marrow derived macrophages are here selected over Raw264.7 macrophage cell line or other cell types (e.g. fibroblasts) as they are the prevalent mouse model and system for studying infection and immune response pathways within the Division of Pathway Medicine, allowing straightforward use of existing experiment data for validation purposes. Treatment options are limited to IFN-γ as this activates a different signalling pathway from other interferons (sections 1.1 and 6.1) and would not be expected to have identical outcomes to those. Although some cross-talk between type II and type I interferon signalling is assumed, this would be better identified by comparing separate meta-analysis. It is unclear how effective particular doses or treatment regimes of IFN-γ are at activating macrophages, prompting the inclusion of any dose and treatment time point used in a particular study's experiment design (with the assumption that they study has been designed within biologically viable parameters). The central thesis hypothesis therefore pertains to common activation pathways of IFN-γ irrespective of dose and treatment regime.

[13] Available in supplementary material, folder "List and description of searches for microarray data sets", file "InterferonMacrophageMicroarrayDatasets.xls"

2. Study contains at least one group of samples of the same type as for criterion 1, but without any treatment intervention.

For RNA expression studies, the fold change between treated samples and reference samples (with all factors except the treatment being identical) is required for determining relative biological effects. Reference or control samples are not always treatment-negative and some studies may for example compare IFN-γ with IFN-β, where this is valid but a different study hypothesis. To study the relative effects of IFN-γ mouse macrophage activation compared to the absence of IFN-γ signal, it is necessary to identify studies that contain the same type of treatment-negative controls as part of their experiment design.

3. All samples in groups described by criterions 1 and 2 are derived from wild-type mice.

In order to pursue different biological hypotheses, biologists often use mouse models that do not express particular genes ("knock-out" mice), that have reduced expression for particular genes ("knock-down" mice) or that have additional genes inserted ("knock-in" mice, transgenic mice) for expression. These changes are more specifically directed than those for mice bred (inbreeds or hybrids) towards particular and stable phenotypes (like C57Bl/6 or Balb/c), and although they can often be characterised phenotypically, it is less clear exactly how these changes may interact with a particular signalling pathway. Given the wide variety of such gene-engineered mice and unknown effects on IFN-γ signalling in macrophages, this inclusion criterion therefore limits studies to those using wild-type mice, albeit from different mouse strains.

Exclusion criteria:

1. Studies not using an Affymetrix gene transcription chip platform

This exclusion criterion is based on the added complexity in mapping and interpreting gene probes between fundamentally different RNA expression platforms. Amongst many other types, these encompass photolithographic nucleotide in-situ assembly with Affymetrix arrays, random bead assembly with Illumina bead arrays, two-channel hybridisations or single-channel hybridisations, and different gene probe designs (e.g. length, evidence-base,

algorithm). While the biological question at the core of this thesis could be better asked by including information from further (otherwise suitable) studies, combining data across multiple microarray platforms is in itself sufficiently complex to warrant separate bioinformatical research.

2. Studies where the only IFN-γ treatment condition is a combined or co-treatment

In macrophages, IFN-γ treatment is considered the priming signal and LPS (lipopolysaccharide) the additional trigger for full activation. Both the priming and triggering of macrophages is biologically highly relevant, but for this thesis the central question relates to the former, as this can build on and be validated by experiment data existing within the Division of Pathway Medicine. As priming and triggering result in a different state for the macrophage, a meta-analysis combining both primed and primed plus triggered macrophages could be carried out but would require rephrasing of the overall hypothesis. Other co-treatments exist, an example of which would be *1alpha,25-DihydroxyvitaminD3. This has* been implicated in mediating immunosuppressive responses (counteracting IFN-γ), and as such this and other macrophage treatments relate to entirely different biological questions and are therefore here excluded from consideration.

Applying the criteria above to the set of 30 studies results in 6 eligible studies, described in summary and individual detail below. Study numbers 1 to 6 are used throughout this thesis and refer to the below table headings.

Table 2.1 Overview of all microarray studies used in thesis

Study	Platform	Organism and background	Number of control samples	Number of IFN-γ samples	Treatment details
1	MGU74av2	Mouse (C57B/6)	3	3	10 u/ml for 24 hours
2	MGU74av2	Mouse (Balb/c)	3	2	100 u/ml for 6 hours or 48 hours
3	MGU74av2	Mouse (Balb/c)	2	2	100 u/ml for 6 hours
4	MGU74av2	Mouse (C57B/6)	2	1	10000 u/ml for 6 hours
5	Mouse430A_2	Mouse (C57B/6)	1	1	500 u/ml for 24-48 hours
6	Mouse430A_2	Mouse (C57B/6 and Balb/c)	7	7	50 u/ml overnight

Table 2.2 Description of study 1

Study ID	1
Study subject	Effect of interferon and/or murine cytomegalovirus infection on macrophage gene expression
Publication	(Kropp, Robertson et al. 2011)
Data source	GPX-MEA: GPX-000029.1
Organism	Mouse (C57B/6 background)
Cells	Macrophages differentiated from bone marrow cells of the femora and tibia of 10-12 week old mice, cultured in medium for 8 days (with interventions as listed below)
Treatment	Recombinant murine IFN-γ at a concentration of 10 u/ml. Control samples were in culture for 8 days, IFN-γ treatment was applied to culture on day 7 for 24 hours.
Array platform	Affymetrix MGU74av2 (contains ~6000 functionally characterised genes and ~6000 expressed sequence tags)
Sample sizes	3 control samples, 3 IFN-γ treated samples. Biologically independent replicates, i.e. individual mice.
Data reference	See Excel spreadsheet for TFID 23 or GPX identifier
Notes	There is uncertainty in the timing of treatment, based on the database, the cited paper and DPM scientists, meaning the treatment may have occurred on day 7 ± 1, and there may be total duration of IFN-γ treatment of 24 hours or 48 hours. This data set should be publicly available, but at time of writing database settings did not allow for external access.

Table 2.3 Description of study 2

Study ID	2
Study subject	IFN-γ induced gene expression in bone marrow derived macrophages following mock, UV treated MCMV and MCMV infection
Publication	(Popkin, Watson et al. 2003)
Data source	GPX-MEA: GPX-000034.1
Organism	Mouse (probably Balb/c background)
Cells	Bone-marrow derived macrophages, differentiated for 7 days in culture prior to treatment
Treatment	100 u/ml IFN-γ, one sample for 6 hours, one sample for 48 hours
Array platform	Affymetrix MGU74av2 (contains ~6000 functionally characterised genes and ~6000 expressed sequence tags)
Sample sizes	3 control samples, 2 IFN-γ treated samples. Biologically independent replicates, i.e. individual mice. Controls match time points of treated samples.
Data reference	See Excel spreadsheet for TFID 24
Notes	Same publication source and author as study #3, but a different data set. Mouse background is not mentioned in this paper, but in the protocol of cited journal paper (reference 10). Detail on culturing BMDM is also not given in this paper, and the cited sources for detail do not actually contain exact details.

Table 2.4 Description of study 3

Study ID	3
Study subject	IFN-γ induced gene expression in CIITA -/-, IRF1-/- and control B6 bone marrow derived macrophages
Publication	(Popkin, Watson et al. 2003)
Data source	GPX-MEA: GPX-000035.1
Organism	Mouse (probably Balb/c background)
Cells	Bone-marrow derived macrophages, differentiated for 7 days in culture prior to treatment
Treatment	100 u/ml IFN-γ for 6 hours
Array platform	Affymetrix MGU74av2 (contains ~6000 functionally characterised genes and ~6000 expressed sequence tags)
Sample sizes	2 control samples, 2 IFN-γ treated samples. Biologically independent replicates, i.e. individual mice.
Data reference	See Excel spreadsheet for TFID 25
Notes	Mouse background is not mentioned in this paper, but in the protocol of cited journal paper (reference 10). Detail on culturing BMDM is also not given in this paper, and the cited sources for detail do not actually contain exact details.

Table 2.5 Description of study 4

Study ID	4
Study subject	Contribution of Interferon-b to the Murine Macrophage Response to the Toll-like Receptor 4 Agonist, Lipopolysaccharide
Publication	(Thomas, Galligan et al. 2006)
Data source	Email correspondence with Stephanie Vogel
Organism	Mouse (C57Bl/6J background)
Cells	Macrophages differentiated from bone marrow cells of the femora and tibia (mouse age not specified), cultured in medium for 6 days.
Treatment	IFN-γ (10,000 u/ml) from a "commercial source", 6 hour duration of treatment.
Array platform	Affymetrix MGU74av2 (contains ~6000 functionally characterised genes and ~6000 expressed sequence tags)
Sample sizes	2 control samples, 1 IFN-γ treated sample. Biologically independent replicates, i.e. individual mice.
Data reference	See Excel spreadsheet for TFID 30
Notes	The publication cited above only provides subsets of data and the data used here have not been submitted to an online repository and are not addressed in that publication because a final decision was made by the authors to focus the manuscript on interferon beta treatments only.

Table 2.6 Description of study 5

Study ID	5
Study subject	Investigation of molecular mechanisms of macrophage activation through IFN-γ and 1-alpha,25(OH)2D3 in wild-type and Vdr (Vitamin D receptor) knock-out mice
Publication	(Helming, Bose et al. 2005)
Data source	ArrayExpress: E-GEOD-2421
Organism	Mouse (C57BL/6 background)
Cells	Macrophages differentiated from bone marrow cells of the femora and tibia of 8-14 week old mice, cultured in medium for 7 days
Treatment	Recombinant murine IFN-γ, diluted in phosphate-buffered saline (PBS) to a final concentration of 500 u/ml. Macrophages were cultured in medium in the presence or absence of IFN-gamma for 24-48 hours
Array platform	Affymetrix Mouse430A_2 (based on Affymetrix documentation, contains ~14000 well-characterised mouse genes)
Sample sizes	1 control sample, 1 IFN-γ treated sample. Both from individual mice.
Data reference	See Excel spreadsheet for TFID 12 or ArrayExpress identifier

Table 2.7 Description of study 6

Study ID	6
Study subject	Effect of macrophage infection with Yersinia enterocolitica with respect to different strains and host resistance
Publication	(van Erp, Dach et al. 2006)
Data source	ArrayExpress: E-GEOD-2973
Organism	Mouse (C57B/6 and Balb/c background)
Cells	Macrophages differentiated from bone marrow cells of the femora tibia, cultured in medium for 8 days
Treatment	IFN-γ (presumably recombinant murine, but this is not detailed) at a concentration of 50 u/ml, macrophages treated overnight.
Array platform	Affymetrix Mouse430A_2 (based on Affymetrix documentation, contains ~14000 well-characterised mouse genes)
Sample sizes	7 control samples (4 Balb/c, 3 C57Bl/6 mice), 7 (4 Balb/c, 3 C57Bl/6 mice) IFN-γ treated samples. Biologically independent replicates, i.e. individual mice.
Data reference	See Excel spreadsheet for TFID 13 or ArrayExpress identifier
Notes	Online repository contains only processed data, unprocessed Affymetrix CEL files were requested and obtained by email

2.1.3 Combining microarray data sets

Microarray platforms are subject to different design in terms of location of probes on arrays, which genes are contained as a probe or probes on an array, how many probes represent a single gene and how many times a given gene is represented. In order to perform a meta-analysis of several studies, it is necessary to ensure that the identity of measured genes is the same for all microarray platforms used. It is possible to (at some risk) simply match data sets on official gene symbols, or to match them on any available unique ID (e.g. EntrezGene). However, as all microarray studies used in this meta-analysis use one of two different Affymetrix chips, it is possible to use an Affymetrix-provided algorithm to match gene probes across arrays based on the amount of overlap in their actual nucleotide sequence.

After normalisation and summarising of probe sets to gene level, Affymetrix chip MOE430A2 contains 22690 gene probes. Chip MGU74AV2 contains 12488 gene probes. Since Affymetrix identifiers changed between these two chip types, a matching based on Affymetrix probe identifiers results in only 43 matches, necessitating the method introduced above and detailed below.

Affymetrix provide product comparison spreadsheets[14] based on the alignment of target gene sequences in two different platforms. In this case, the chosen comparison file[15] contains 14230 "good match" gene probes that are contained on both array platforms, where "good match" is defined as the two sequences from different chip platforms matching all of the following criteria: their sequence overlap is > 90%, the minimum sequence length of the shorter of the two representative[16] sequences is greater than 100 bases (that is, more than 100 bases can overlap between these representative sequences, this number does not relate to the actual array-bound sequences chosen from these representative sequences), and the number of Perfect Match probes in either sequence aligning with the other sequence is greater than 1.

Affymetrix microarray platforms often use more than one chip to represent a full genome, usually referred to as "A" chip and "B" chip. The studies used for meta-analysis in this thesis make use of only the "A" chips, which means the total of 14230 matched gene probes includes those on "B" chips. Consequently, once these are removed, a total of 9812 gene probes remain for meta-analysis based on the two "A" chips used in all studies.

2.2 Software tools and resources

Microarray data were processed and analysed using R (R Development Core Team 2009), available from http://www.r-project.org. The R Versions used for the implementation of the body of work relevant to this thesis were 2.8.x and 2.9.x (x denoting subversions). All R scripts relating to this thesis are provided as electronic supplementary material. Apart from simple utility functions or functions already provided with R, several packages from the Bioconductor (Gentleman, Carey et al. 2004) repository were also used. These are listed in the following table:

---

[14] http://www.affymetrix.com/support/technical/comparison_spreadsheets.affx?pnl=1_2#1_2

[15] mgu74av2_vs_mouse430_good_match.zip

[16] The 'representative sequence' is the sequence used at the time of chip design to represent the transcript that the probe set on the Affymetrix chip measures. Individual probe sequences (and the probe sets they constitute) on the array are shorter and selected from a ~600 base region at the 3' end of the representative sequence.

Table 2.8 Existing R/Bioconductor packages used

Package name	Version	Purpose
affy	1.20.2	Loading and processing of Affymetrix microarray (CEL) files (Gautier, Cope et al. 2004)
geneplotter	1.20.0	Graphics functions (Gentleman)
arrayQualityMetrics	1.81	Quality control metrics and graphs (Kauffmann, Gentleman et al. 2009)
gplots	2.6.0	Advanced graphs, specifically heat map variations (Martin 2011)
marray	1.20.0	Colour palettes (Wang, Nygaard et al. 2002)
RankProd	2.14.0	Rank Product analysis (Breitling, Armengaud et al. 2004)
GeneMeta	1.14.0	Effect size *d (Lusa)*
Epi	1.0.8	ROC statistics (sensitivity and specificity of results) (Carstensen 2012)
preprocessCore	1.4.0	Quantile normalisation (only used in chapter 4, imputation) (Bolstad)
impute	1.14.0	Missing value imputation functions (linear regression, KNN)
pcaMethods	1.18.0	Missing value imputation (BPCA) (Stacklies, Redestig et al. 2007)
GEOquery	2.6.0	Retrieve gene expression data sets from GEO database (Sean and Meltzer 2007)

For bioinformatical interpretation, statistically significant results were saved as lists and imported into Ingenuity Pathway Analyzer (IPA) where they were assessed for relevance to particular gene ontology function and transcription pathways/networks. Individual genes of interest were also assessed against a number of other resources containing either consensus or computationally acquired knowledge or experimental data. These are listed in table 2.9.

Table 2.9 List of bioinformatics resources

Resource	Availability and citation where provided	Purpose
Ingenuity® IPA	www.ingenuity.com	Gene enrichment analysis for networks, pathways, gene ontology
NCBI	http://www.ncbi.nlm.nih.gov/ (Maglott, Ostell et al. 2011)	Gene annotation and literature searches
NEXTBIO	www.nextbio.com (Kupershmidt, Su et al. 2010)	Known association of individual genes with other studies, diseases, pharmacology, experiments
BioGraph	www.biograph.be (Liekens, De Knijf et al. 2011)	Known relation of individual genes with other genes in established networks/pathways
GeneCards	www.genecards.org (Safran, Dalah et al. 2010)	Compiled annotation on individual genes
WikiGenes	www.wikigenes.org (Hoffmann 2008)	Compiled annotation on individual genes
iHOP	www.ihop-net.org (Hoffmann and Valencia 2004)	Gene-gene relations via protein knowledge.
GEViSE	http://oriol.gti.ed.ac.uk:8380/GEViSE/	Gene expression profile browser for studies conducted within DPM, or relevant to DPM's research.
yEd	www.yworks.com	Graph editor for drawing pathways or general flowcharts
Cytoscape	www.cytoscape.org (Cline, Smoot et al. 2007)	Drawing and analysis of network graphs
Interferome	www.interferome.org (Samarajiwa, Forster et al. 2009)	Online database of interferon stimulated genes
STRING	string-db.org (Franceschini, Szklarczyk et al. 2013)	Protein-protein interaction networks
oPOSSUM	opossum.cisreg.ca (Sui, Fulton et al. 2007)	Transcription factor binding site analysis

2.2.1 Imputation of missing values

Imputation of missing values in context of meta-analysis employs several R functions, which are employed within a more complicated wrapper in order to apply them to resampled data sets and across multiple microarray studies. As such, the code is too long to reproduce in print, but the complete R code for imputation of missing values and their assessment is provided in electronic form with this thesis ("ImputationMissingValues_HF.R").
The imputation algorithms are described in more detail in chapter 4, this section lists the methods and the corresponding R code relevant for its implementation.

K-Nearest-Neighbor or KNN (Troyanskaya, Cantor et al. 2001) uses R function impute.knn(…,k=30, rowmax=0.9) in package impute v1.14 as provided. For array-wise imputation, the input data matrix is rotated through 90 degrees (and arguments k set to 5 and colmax set to 0.8), so that similar samples rather than similar genes are used to impute missing samples.

Bayesian Principal Components Algorithm or BPCA (Oba, Sato et al. 2003) uses R function pca(…,method="bpca", nPcs=6) in package pcaMethods v1.18.0 as provided.

Imputation by linear regression uses a generic lm() R function for linear models in the basic stats package, using non-missing gene expression values from all other variables (in this case, those variables are samples or arrays) as explanatory variables. An alternative version is also employed where the explanatory variables are only those samples from the same biological treatment group (control or interferon gamma treated) as the missing sample.

Imputation by ranked sets uses Euclidean distance between genes to identify genes similar to the one with missing values in another study. The 50 genes closest in their expression level profile are then identified by forming the rank product, which is 1/n times the product of the ranks of n Euclidean distances, where n is the number of studies the gene in question is not missing in. For each of these 50

genes in each of the studies (in which the gene is not missing in), inverse normalised weights ((1/distance) / sum(1/distance)) are obtained based on the Euclidean distance, then averaged for each gene across these studies. Finally, a weighted mean (using average weights above) is computed from these 50 genes in the study where there the gene in question is missing and used as the imputation estimate.

2.3 Microarray target sample and array processing

The steps described in this section are not immediately relevant to replicating the efforts contained within this thesis, as they precede the processing and analysis of numerical data. However, the standard laboratory procedures for using Affymetrix arrays are outlined here for better understanding of their place within the full laboratory-to-results workflow. These steps are described in detail in the official Affymetrix manual[17] and a schematic overview[18] is presented in figure 1.2 (chapter 1). They include all processes from amplification of the initially extracted total RNA sample to biotin-labelled cRNA ready to be hybridised to a microarray.

After hybridisation is complete, microarrays are scanned in an Affymetrix GeneChip instrument, where a confocal laser scans each microarray and obtains an image where each individual probe's expression level is represented by fluorescent intensity. Using Affymetrix software (GCOS or GeneChip Operating Software prior to 2008, AGCC or Affymetrix GeneChip Operating Software from 2008), an image quantification algorithm converts the images (DAT files) to numerical signal intensity levels (CEL files).

[17] Eukaryotic Sample and Array Processing (http://media.affymetrix.com/support/downloads/manuals/expression_s2_manual.pdf)
[18] Figure 6, Dalma-Weiszhausz, D. D., J. Warrington, E. Y. Tanimoto and G. Miyada (2006). "The affymetrix GeneChip (R) platform: An overview." DNA Microarrays Part A: Array Platforms and Wet-Bench Protocols **410**: 3-28.

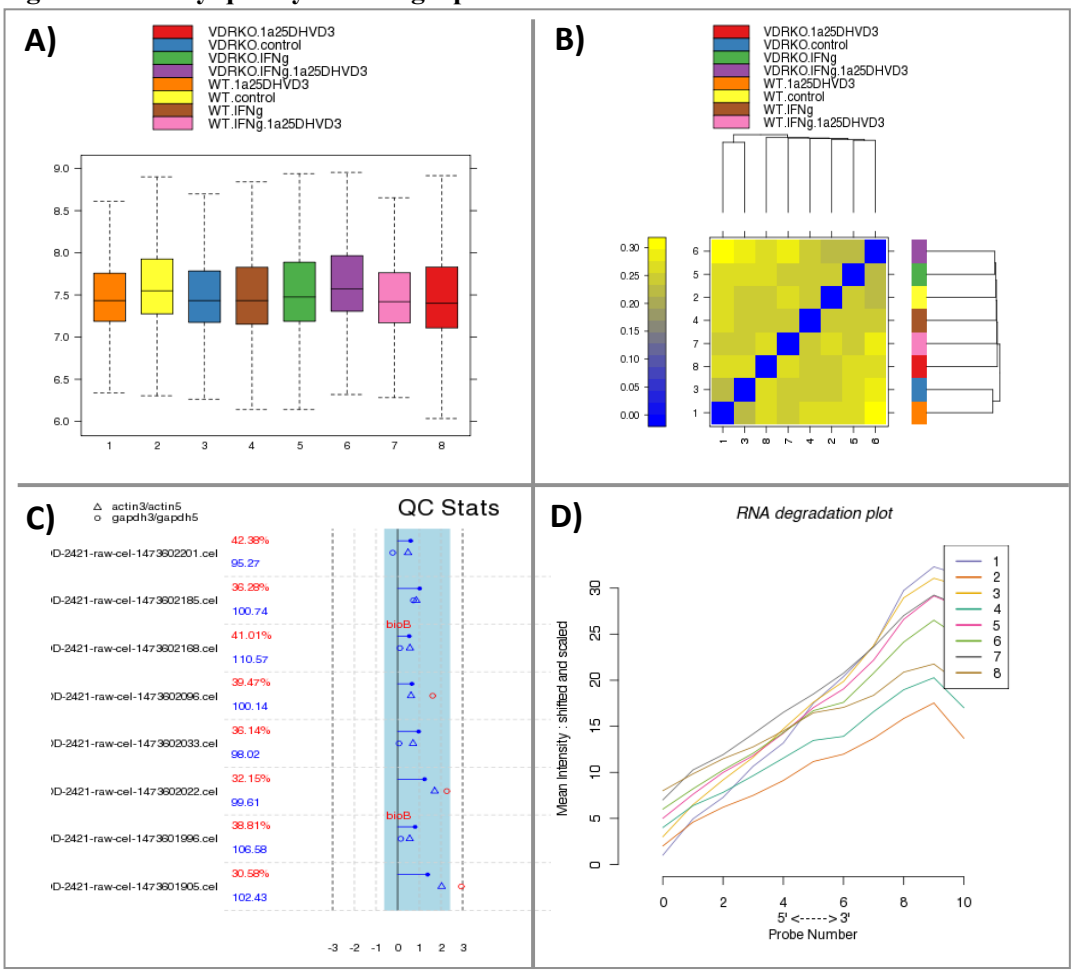
2.4 Data quality control

Data quality of each array in each study is inspected through graphs and metrics generated by Bioconductor package arrayQualityMetrics (Kauffmann, Gentleman et al. 2009). This involves the performance of specific control probes on the array, the per-array data distribution, and similarity across arrays and numerical representation of the array surface. Individual arrays are excluded from further processing and analysis if a majority of the aggregated evidence indicates low quality. For each study, a quality control report ("Qmreport.html") is included in the supplementary material[19]. A selection of these graphs is shown in figure 2.1 below and quality control assessments are described in more detail in the following sections. A complete description of the output the arrayQualityMetrics package produces is available as a tutorial[20] from EMBL-EBI (Gabriella Rustici, Audrey Kauffmann 2009).

[19] Included as electronic files in folder "Study Quality Control Reports"

[20] http://www.sbforum.org/cmsimages/presentations/BioC_QM%20Tutorial.pdf (included here under Creative Commons Attribution-Share Alike 3.0 license: http://creativecommons.org/licenses/by-sa/3.0/)

Figure 2.1 Array quality control graphs



Shown are data distribution box-and-whisker plots for all the gene probes on an array (**A**) prior to data processing, with X-axis indicating an array and Y-axis representing $\log_2$ scale signal intensity. Array groupings due to biological factors or unintended batch effects are identified through a heatmap representing distance between all data points of any two arrays (**B**). This distance is computed as the mean absolute difference between all unprocessed gene probe signal intensities and then represented by a colour (0 = blue = identical array expression levels). A clustering algorithm arranges (dendrograms) arrays in this heatmap by those distance estimates. A summary report (**C**) provides information about expected and observed Affymetrix control probe signal intensities, array background and array scaling factors. Red plotting symbols indicate abnormal values and a problem with the array. The RNA degradation plot (D) estimates each array's RNA quality by averaging signal at a given probe position (from 5' to 3' end) across all probes on the array, and concern is raised if any arrays are very dissimilar to others.

Control probes. This method serves as a proxy measurement to determine RNA quality issues on an array. Affymetrix chips contain sets of control probes that are designed to provide information about the integrity of hybridised RNA. The chip platforms used in this thesis contain GAPDH and Actin probes from both the 3'

and 5' end of the transcripts. The 3'/5' ratio of their measured intensity values is expected to remain below 3, although this threshold is arbitrary for any given organism or sample type, and the ratio is also expected to be smaller for the shorter transcript (GAPDH). This ratio is therefore mainly used to identify individual deviating arrays in a study[21].

Per-array data distribution. This method is based on the assumption that a biological intervention (in this case, various treatments applied to macrophages) should affect only a small proportion (less than 5-10%) of the gene probes on an array. Therefore, data distributions should be almost identical for all samples. For each microarray sample in a given study, the data distribution of (log-transformed) signal intensity values across all gene probes is assessed through probability density plots and box-and-whisker plots. If probability density plots and box-and-whisker plots show individual samples of deviating frequency distribution, then this is an indication for quality issues. Shifted but otherwise identical distributions are not included in this statement, as this may be due to systematic factors equally affecting all gene probes, and this can be addressed through data normalisation as explained below.

RNA degradation. Also known as RNA digestion plot, this assessment does not make use of selected control probes to estimate quality of RNA on the array, instead it uses all probes on the array. This is made possible by the way Affymetrix chips measure gene expression, i.e. for each gene there are a number of probes (11 for many Affymetrix chips) representing it, each covering a different part of the sequence of the target gene. For each of the 11 probes covering a distinct target sequence region between the 5' and 3' ends, average signal intensity is calculated across all target genes and plotted (shifted and scaled to allow a direct comparison between arrays). The resulting RNA profiles are expected to start lowest at the 5' end (where degradation begins) and rise for 3' end probes. In the absence of definitions for quality thresholds, main purpose of these graphs is to compare slopes for all arrays and identify individual outlier arrays, which will have more extreme slopes than other arrays.

---

[21] Source: Affymetrix (http://www.affymetrix.com/support/help/faqs/ge_assays/faq_17.jsp)

Multi-array metrics. Based on the assumption that only a minority of genes (out of the entire genome represented on the array) should be affected by the biological intervention made, there should be a good overall consistency of signal intensity across all samples. In contrast to per-array data distributions, the methods listed here retain gene identity information across samples. Between-array consistency is assessed by a) a heat map representing all pairwise sample distances (or conversely, correlation), b) MA-plots showing the log-scaled intensity levels (X-axis, referred to as "A") and log-scaled differential expression (Y-axis, referred to as "M") between each individual array and a virtual array representing an average across all arrays, and c) probe-level estimates for each probe (individual sequence deposited on array, not at gene level) across all arrays in form of NUSE (per-gene standard error in relation to median standard error across all arrays) and RLE (per-gene expression level in relation to median expression level across all arrays).
Physical chip problems. This method aims at identifying large scratches or chip regions where large numbers of probes have failed to work for reasons associated with the physical chip structure or the hybridisation/washing procedures. Each array has the rank (indicated by colour) of all individual probes' signal intensity plotted by its array coordinate location, visualising the numerical signal levels (which may differ from the original scanned images in that those do not aggregate pixel information).

2.5 Microarray data processing

Background correction, probe set summation and between-array normalisation follow the Affymetrix-specific methodology referred to as RMA or Robust Multichip Average (Irizarry, Hobbs et al. 2003). Arrays are not independently processed, but in context of all arrays in a study, which requires that no arrays be discarded from a study prior to meta-analysis as discussed in this thesis. Within the algorithm, three separate stages are performed in sequence, as detailed below.

Background correction is based on the assumption that signal intensity measured for Affymetrix "Perfect Match" or PM (probes representing a gene of interest)

probes is comprised of two independent components, one being real binding events during hybridisation (following an exponential distribution when considering all probes on an array) and the other being non-specific binding or cross hybridisation and optical noise (following a normal distribution and truncated at 0). This means the expected true expression value – given the observed signal intensity – can be obtained by adjusting or rather solving for the observed signal intensity with the mean and variance parameters (normal distribution) and the rate parameter (exponential distribution), all estimated from probes on array. It should be noted that RMA background correction ignores mismatch (MM) probes that are included on Affymetrix arrays as a hybridisation control for each given PM probe. Other Affymetrix data processing algorithms may use these to subtract from their corresponding PM probes. Irizarry et al have found this to introduce noise, and instead follow the above strategy. Specifically, per-array observed PM raw signal scale intensity S is modelled as $S=X+Y$. X in this case is the probe hybridisation signal and follows an exponential distribution with its parameter α. Y is the background noise component and follows a normal distribution with its parameters μ and σ^2. The data transformation using these two distributions then is solved as:

$E(X|S) = a + b[\phi(a/b) - \phi((s-a)/b)] / [\Phi(a/b) - \Phi((s-a)/b)-1]$,

where $a = s - \mu - \sigma^2\alpha$; $b = \sigma$; ϕ = normal density; Φ = cumulative density.

Normalisation of data across samples is performed with background-corrected data still on the linear scale and at the level of individual (Perfect Match) probes. All normalisation methodologies are targeted at making whole-array data distributions comparable across multiple arrays by accounting for systematic effects on signal intensity levels, i.e. any factors that are expected to affect every probe on an array to the same degree. This may for example include hybridisation efficiencies, laser scanner settings, or biological sample processing. The approach is inappropriate if the biological condition on some arrays is expected to affect more than 5-10% of the probes on an array. Normalisation within the RMA framework forces whole-array data distributions to be identical for all arrays.

When applied, the algorithm sorts all probes (not aggregated to gene level) on each array by measured signal intensity. Each arrays' lowest quantile (probe with lowest signal value) is set to be the average of all lowest quantiles across all arrays. This is repeated for every quantile (probe). When all probes' intensity values are so adjusted, the full data set of all arrays is returned to its original probe ordering.

Probe summarisation is the process of combining the signal of all probes on an array representing a given target gene. Given different expected affinities of probes, a simple average across a probe-set in one array is rejected in favour of a robust median polish taking account of a probe-sets signal intensities in all arrays. Within the RMA algorithm, all perfect match probe intensities are first $\log_2$-transformed, allowing any error to have constant variance (because signal error is expected to be proportional to measured intensity on the original scale). Tukey's median polish procedure is applied to the $\log_2$-transformed data, with each probe-set represented as a matrix of rows and columns, where the rows are arrays and the columns are individual probes of a probe-set. Row and column medians are iteratively subtracted from row and column values until convergence (all row and column medians = 0) determines a set of final residual values. Subtracting these residuals from the original matrix for a given probe set allows per-array estimation (arithmetic mean of those fitted values per row) of the expression level of the given gene in a given array.

Following RMA, probe signal intensity values are now background-corrected, quantile-normalised, $\log_2$ transformed and summarised measures of the expression level of a given gene in a given sample.

2.6 Statistical analysis of individual microarray studies

For each gene in each study, the null hypothesis of "no differential expression between IFN-γ treated samples and control samples" is tested. While the presence of biological samples not relevant to the meta-analysis was necessary for initial

data processing, these are no longer required and removed from further analysis, resulting in the per-group sample sizes shown in the study summary tables in section 2.1.2. For regular analyses of individual studies, a single statistical model is usually employed, but the use of three different meta-analysis models requires multiple statistical tests corresponding to each meta-analysis model. All statistical tests described here take as input only the microarray samples matching the two biological conditions relevant to this thesis: untreated controls and IFN-γ treated samples. For all tests, the input data are pre-processed and normalised, including transformation of expression values to $\log_2$ scale because gene expression values are assumed to follow a log-normal distribution (Quackenbush 2002). For the non-parametric Rank Product test, data transformation is inconsequential.

Rank Product test (Breitling, Armengaud et al. 2004). This test is the single-study analysis equivalent of the Rank Product meta-analysis. However, the results of this test (p-values) are also used as alternative input for Fisher's meta-analysis method as described in chapter 3. Using the RankProd package for R, statistically significant differentially expressed genes are identified based on calculating each gene's rank product statistic, which is the product of all possible pairwise-sample ranked differential expression values K (difference on logged data) between the control and the IFN-γ condition, to the power of $1/K$. Statistical significance is assigned by measuring how extreme the statistic is when compared to a Null distribution of possible Rank Product statistics based on permuting the original data set gene-wise B times (in this case, B=100) and computing the Rank Product statistic each time. The approach is repeated once more, with ranking of differential expression reversed (i.e. identifying down-regulation with respect to control condition). This test uses ranking across all genes measured in a sample instead of testing each gene in isolation, and therefore includes more information than other statistical tests. The end result is a list of genes with a statistical significance p determining how consistently high (or low, in the reverse case) each gene is ranked in terms of expression ratios between control group samples and IFN-γ group samples. The specific R function used is RP(), with 100 permutation runs.

Welch's t test (Welch 1947). This test calculates the per-study significance of the difference in mean (log_2-transformed) expression between control and IFN-γ samples, the per-study p-values so obtained are used as input for Fisher's meta-analysis method. Welch's t test is almost identical to a standard Student's t test and estimates the statistical significance of the difference in the means (per gene probe, in this case) between two groups of observed expression values, given their variance. As opposed to the Student's t test, it replaces an estimate for pooled group variance with individual estimates for per-group variance, therefore allowing each group in the comparison to have a different variance. Genes are tested one at a time and the end result is a list of genes with statistical significance assigned to their mean difference in expression between control and IFN-γ group. The specific R function used for this is ttest().

Effect size z score. This is the per-study test statistic equivalent to the combined effect size used for meta-analysis. Apart from not combining the effect size estimates across studies (as is done in a meta-analysis), the equations used here are identical to those discussed in detail in chapter 3 (equations cited here as appropriate). Effect size *d* (eq. 3.10) is very similar to a standard Student's t statistic, but instead of standardising the difference in group means to the pooled standard error, it standardises the difference in group means to the pooled standard deviation (the denominator), i.e. it does not explicitly account for sample size and is therefore biased. However, subsequent to calculating the biased effect size statistic, a sample size bias correction (eq. 3.11) is applied and the variance of this unbiased effect size estimator estimated (eq. 3.12). A z statistic is then constructed by standardising the per-study unbiased effect size to the per-study standard deviation of the effect size (eq. 3.19, but for individual study estimates, not combined estimates across studies). This z statistic follows a standard normal distribution and its statistical significance can therefore be obtained from the standard normal reference distribution table, in this case specifically using the pnorm() function in R with the mean and standard deviation parameters set to 0 and 1, respectively and in accordance with a standard normal distribution.

2.7 IFN-γ pathway reference lists

For chapter 3, and in order to quantify relevance of meta-analysis results against known IFN-γ and immune response gene transcription pathways, current consensus information is obtained from 2 alternative sources. The first is NCBI Gene database queries, with four different sets of search criteria and limited to Mus musculus. These are referred throughout as gold standard or reference lists A to D in this thesis. These are not curated for scientific trustworthiness and simply identify all genes ostensibly connected with type I or type II interferon pathways. In contrast to this, the second type of list is manually curated and referred to as the DPM gold standard or reference list. All lists are included in the supplementary material[22], with the DPM list also included as Appendix A12.
The lists described in this section are used for a quantitative assessment of meta-analysis results and are not primarily based on microarray gene expression data. This differs from the approach taken in the biological assessment of meta-analysis results, which uses microarray-derived reference sources (section 2.8)

2.7.1 NCBI gene reference lists

In a two-stage process, the NCBI [Gene] database was first queried with relevant search parameters to identify gene transcripts and corresponding EntrezGene IDs, and the retrieved results then matched against the Affymetrix gene probes contained on the two chip platforms used for meta-analysis in this thesis.

Stage 1. The query results used in this thesis were obtained on 14 Jan 2009, annotation updates subsequent to this date are therefore not reflected in those lists. All queries were run through NCBI's Gene portal[23] (Maglott, Ostell et al. 2011) and the results are summarised in table 2.10. The four queries submitted to NCBI

[22] Available in folder "NCBI and DPM reference gene lists", DPM list also included in appendix A12
[23] http://www.ncbi.nlm.nih.gov/gene/

are not independent and investigation of overlap through Venn diagrams identified considerable overlap between all lists.

Stage 2. Using the EntrezGene identifiers obtained in the previous stage, further identifiers and annotation were added by retrieving Ensembl[24] (Flicek, Ahmed et al. 2013) data for Mus musculus from Biomart[25] (Kasprzyk 2011), with explicit reference to the two Affymetrix microarray platforms Moe430a and Mgu74av2. This was implemented through package biomaRt within R. Specifically, the biomaRt package was used to access the dataset "mmusculus_gene_ensembl", version NCBI37. This stage reduces the number of gene transcripts in each list because it discards those not present on a given microarray platform.

Table 2.10 Query-based reference list of known biological pathways

Reference list	Query terms	Retrieved gene transcripts	Matching meta-analysis microarrays
A	"ifng" OR "interferon gamma"	532	190
B	"immune response"[GO]	944	244
C	"immune response"	1218	343
D	"interferon"	1219	271

Queries were submitted to the NCBI[Gene] database. Searches were limited to one organism (Mus musculus). Searches for lists A,C,D were done on database fields containing gene symbol and gene description, list B was limited to the database field (indicated as [GO] in this table) containing gene ontology categories. Letter case has no influence on the search. The number of retrieved gene transcripts is the total of all NCBI gene records where the specified search terms were found in the specified database fields. In contrast, the number of genes matching meta-analysis microarrays is the subset of all retrieved gene transcripts that are contained in the combined (number of genes = 9812) microarray data sets used for meta-analysis.

2.7.2 Division of Pathway Medicine gene reference list

This list is based on extensive manual curation of literature and databases, undertaken by several DPM MSc students and their supervisors over several years with the aim of compiling a consensus signalling pathway map relating to

[24] http://www.ensembl.org/
[25] http://www.biomart.org

macrophages (Raza, McDerment et al. 2010). Although the target signalling pathway is intended to be based on human macrophages, information is compiled from human and mouse, and using other cellular systems. Relevant data were curated from PubMed, Google Scholar, STRING, BIND, KEGG, Reactome, Ingenuity and consist of protein/gene identifiers, interaction types and other annotation. This information was organised in interaction lists and a pathway database, assembled into a signalling pathway map and refined through expert opinion and comparison with experimental data. The outcome comprises many macrophage signalling pathways, but for the purposes of this thesis the list is limited to the IFN-γ related signalling pathway only. In January 2009, this list subset contained 67 genes that are a) known to be related to the IFN-γ transcriptional network and b) contained on both microarray platforms used for this thesis.

2.8 Reference sources for biological validation of meta-analysis results

2.8.1 Interferome

Interferome (Samarajiwa, Forster et al. 2009) –database available at www.interferome.org- is manually curated from publicly available microarray data sets (36 to date) to consolidate type I, type II or type III interferon regulated genes. The interferome study set is heterogeneous with regard to organism, cell/tissue types and type/subtype of interferon.

Gene symbols present in meta-analysis results were converted to ENSEMBL IDs using Biomart[26]. While the database contains ~2000 genes identified as interferon stimulated, it is limited to accept 100 genes at a time for any input list, and meta-analysis results were therefore submitted in batches of the required size. For each gene, the returned information indicates if it has previously been identified as interferon-stimulated (type I, II, III, or any combination thereof).

[26] www.biomart.org

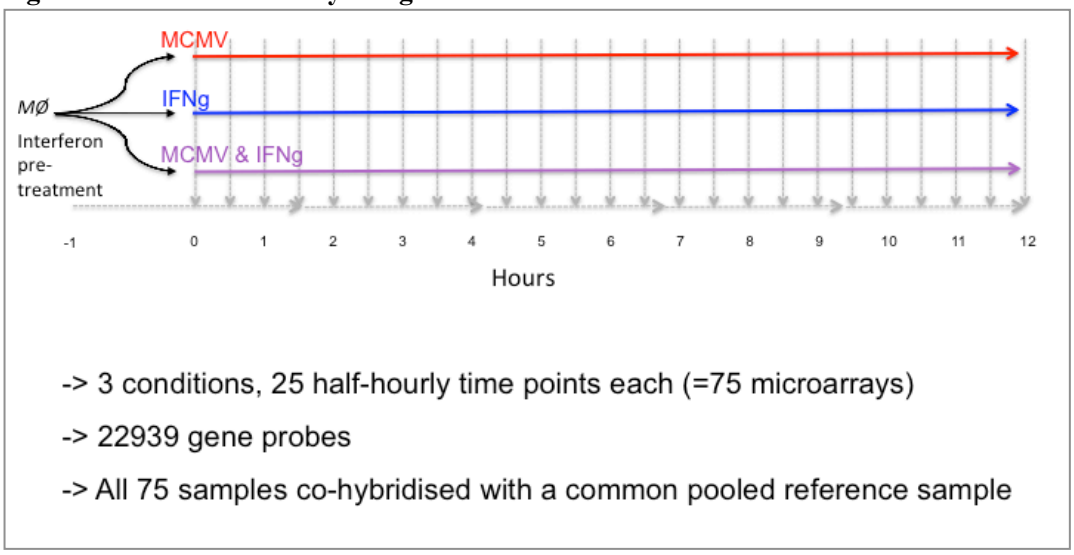
2.8.2 Microarray time course study (MITCH12)

This microarray study consists of 3 biological conditions, murine bone-marrow derived macrophages (BMDM), grown in culture for 7 days, then primed with low-dose Ifn-γ and activated by Ifn-γ, infected with murine cytomegalovirus and both Ifn-γ activated and virus infected. Transcription levels for all murine genes (Agilent arrays) in each condition are measured over 25 half-hourly time points, starting at 0h and ending at 12h. BMDM are extracted from the original set of mice, pooled and then plated out into 75 individual experiments, with each undergoing one of the three specified treatments and an RNA sample obtained at one of the 25 specified time points. Biological inter-subject variation does therefore not contribute to overall variation in this study and the only observable gene transcription differences should be due to the treatment or time of measurement. For purposes of statistical analysis, replicates (BMDMs treated with interferon gamma in more than one experiment per time point and condition) would be preferable in order to assess the experiment noise contributed by effectiveness or duration of treatment, but the final study design sacrificed this in exchange for extended longitudinal observation. Each array is a co-hybridisation of the test sample and a universal reference, where the reference is pooled material from all macrophage cultures across the entire time-course.

Macrophage cultures were established from the bone marrow of ten to twelve week old male Balb/c mice (Charles River Laboratories, Kent, U.K.). Briefly, bone marrow progenitors were flushed from femurs and plated at 8×10^5 cells in 6 well tissue culture dishes (Costar, Corning Inc., NY, USA). The cells were cultured in DMEM/F12 medium supplemented with 10% fetal calf serum (FCS), pen/strep, glutamine and 10% L929 conditioned medium as a source of M-CSF. BMDM were infected with the different viruses at a MOI of 1, Ifn-γ treatment consisted of 10 u/ml (Boelinger Manhaim Corp). RNA was extracted from each sample using Trizol RNA extraction protocol. Mock samples were pooled and labelled with Cy3 while the lyzed or IFN-γ treated samples were labelled with Cy5 using a modification of the Agilent Fluorescent protocol, using half of the standard

Cy3/Cy5 labelled dUTP concentration. The Cy3 labelled pooled control was hybridized with each of the 75 Cy5 labelled samples according to the Agilent Low RNA Input protocol. The dual hybridizations were carried out on Mouse Agilent V2 array (G4121A, 20868 annotated probes), and were scanned on an Agilent Technologies scanner. Agilent feature extraction software (V.A7.5.1) was used to extract numeric data for further analysis.

Figure 2.2 MITCH12 study design



Background noise was corrected for by subtracting background signal from foreground signal, any resulting expression values less than 1 were set to 1. Data were then transformed to $\log_2$ scale. Normalisation between all arrays was performed on the basis of an identified subset of control probes (positive control probe set), separately for test and reference samples. The normalised data set was then filtered to remove "flat" expression profiles, based on performing a ROC analysis on each microarray, using known positive and known negative control probes for identification of a threshold value (an expression level that provides >=80% sensitivity in distinguishing negative from positive control probes). These threshold values are averaged across all 75 arrays (3 time courses with 25 samples each) in the study, resulting in an overall threshold value. In order for a gene probe to remain in the analysis data set, it was required to have an expression level greater than or equal to the overall threshold in 5 or more consecutive time points

for a given time course. The final data set size for analysis (14299 gene probes) consists of all gene probes that match this criterion in at least one of the time courses.

2.8.3 Transcription factor binding site enrichment (oPOSSUM 3.0)

This tool is available at opossum.cisreg.ca (Sui, Fulton et al. 2007).
Searches for Gamma-Activated-Sites (GAS motif) were carried out for two purposes. One was to create a reference list of genes in the mouse genome that contain a GAS motif and are therefore a relevant comparison to microarray meta-analysis of Ifn-γ activated macrophages. A second purpose was to establish which transcription factor binding sites the meta-analysis result lists were enriched for.

In order to use oPOSSUM, gene symbols contained in meta-analysis results were first converted to Ensembl IDs, using the Biomart ID Converter tool available at www.biomart.org. Ensembl IDs are then submitted to oPOSSUM, also specifying the following parameter choices.
Background: all genes (29347) in oPOSSUM database
JASPAR CORE profiles: specific vertebrate profile, select STAT1
TFBS search parameters: default, but upstream/downstream sequence set to 5kb/2kb

Chapter 3

Meta-analysis of the effect of IFN-γ on macrophage gene expression

3.1 Introduction

In this chapter three different and established statistical models and three model variations for meta-analysis are applied to a collection of six[27] microarray gene expression studies that test the effect of IFN-γ stimulation in murine macrophages. The three meta-analysis models comprise Fisher's sum of logs, effect size model and rank product meta-analysis. Subsequently, meta-analysis results are assessed with reference to quantitative metrics both in isolation and with reference to biological expectations. A more detailed and bioinformatic assessment of the biological meaning underlying these results is the focus of chapter 5. Altered variants of the basic meta-analysis models are devised in some cases and also assessed.

The overall goal is to combine the statistical hypotheses tested in individual studies in order to obtain a combined statistical estimate across those studies. The primary questions for this chapter are as follows:

1. In a setting with multiple very small studies, does meta-analysis provide additional results when compared to individual studies?
2. In a setting with multiple very small studies, are there performance differences between the meta-analysis methods here?
3. Can performance-improving alterations to the meta-analysis approaches be identified?
4. In a setting with multiple very small studies, do meta-analysis results show biological relevance?

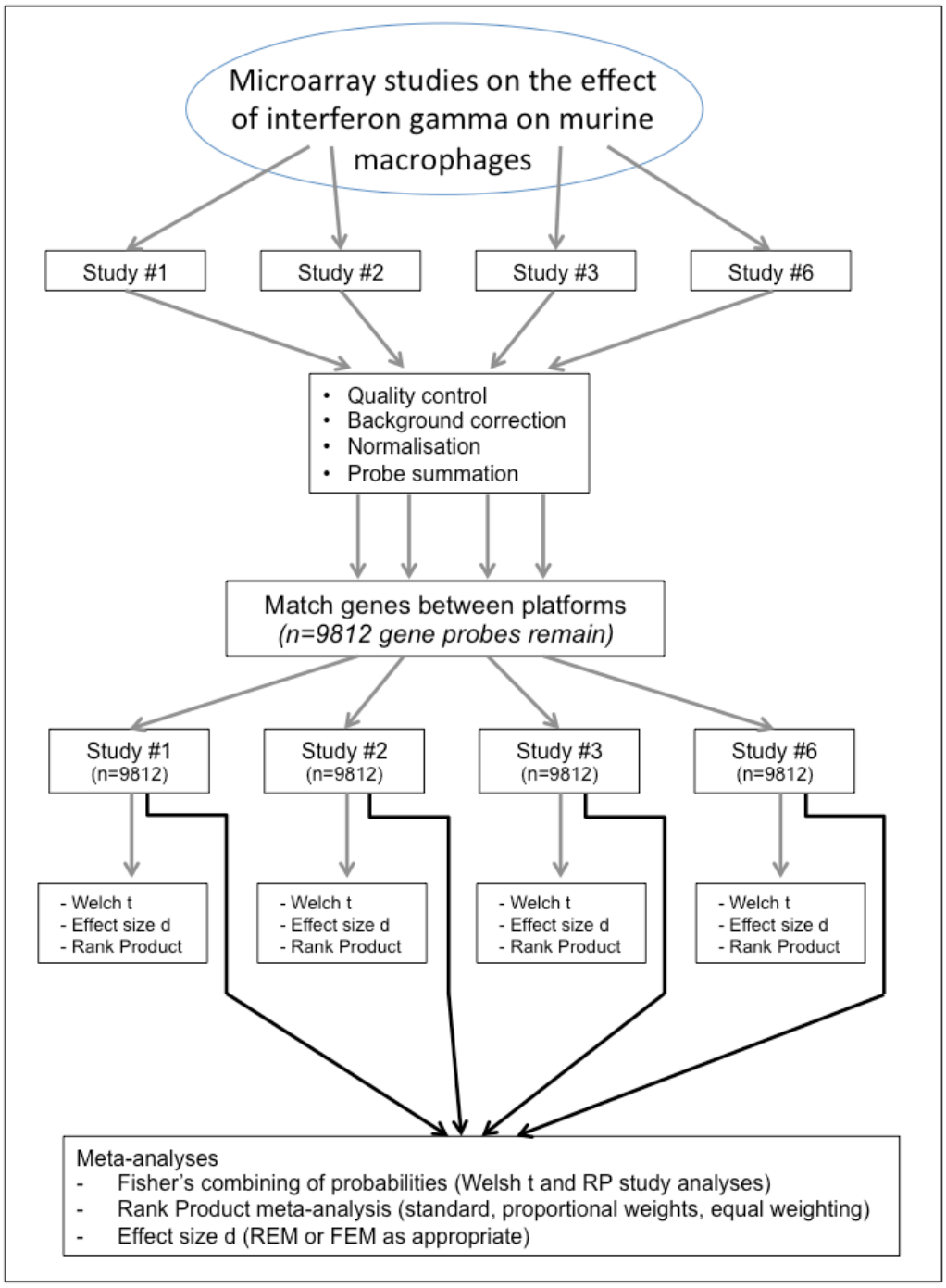
[27] Six studies where the meta-analysis model (Rank Product meta-analysis) allows the inclusion of studies with a single sample in one group, only four studies are methodologically eligible for the other two models.

Although not separately pointed out in every instance of a meta-analysis model or data estimate in this chapter, it is clear that statistical robustness (for both parametric and non-parametric models) will be impacted by the availability of only 4 (6 in some cases) studies, where the sample size of individual studies is also very low. While there are certainly microarray studies with tens to thousands of hybridised samples, these are relatively rare (an estimated 6%, see below). Although meta-analyses are of course being done on these and provide biologically useful results, it is unclear how they perform with the much more frequent smaller and individually underpowered sets of studies. An estimated 33% of experiments have less than 6 hybridisations, where this may include anything from 1 to 6 biological conditions. Reference numbers for this are based on a broadly selected set of all RNA array-based studies on mus musculus, with $\tilde{x} = 11$ (MAD=7.4) assays per experiment (total experiments = 6355 at 21 Jun 2012[28],), across all biological conditions/groups it contains. To provide some admittedly very broadly defined estimates for ostensibly large and small studies, only 6% of all experiments consist of more than 50 hybridised samples, 33% consist of 6 samples or less. The number and size of studies here fits into this scenario and is comparable to or higher than that used e.g. by Breitling or Choi in their respective meta-analysis approaches.

The steps prior to assessment are summarised in a flowchart, figure 3.1.

[28] source: http://www.ebi.ac.uk/arrayexpress/browse.html

Figure 3.1.1 Study selection, analysis and meta-analysis workflow



Not shown in this overview are studies #4 and #5, as they are too small to contribute to parametric methodology since means or variances cannot be computed. However, studies #4 and #5 are included as additional inputs for Rank Product meta-analysis as this best represent the actual problem investigated. R scripts corresponding to this figure are included as electronic supplementary files for particular steps: Quality control, background correction, normalisation and probe summation can be found in "TFIDXX_Processing.R". Mapping genes between array platforms is carried out in "Script5a_CreateMetaset.R". Analysis and meta-analysis of individual studies by a given method is contained in "Script6_MetaSet_XX.R". Additional summary graphs (across meta-analysis models) are added in "Script7_GraphingResults.R".

3.1.1 Study selection

Based on the collection of microarray studies described in chapter 2, the data sets chosen represent the narrow field of studies with overlapping experiment design parameters. Their area of investigation matches biological hypotheses that the Division of Pathway Medicine is both familiar with and has data for, which supports validation and interpretation of results. Inclusion criteria for the six studies used were murine bone marrow derived macrophages, IFN-γ treatment, a control condition and Affymetrix-branded microarrays. As described in detail in chapter 2, this still leaves a broad range of differences between the studies, i.e. interferon dose and timing, mouse model, laboratory protocols, probe-sequence. While this means a compromise in the biological questions that can be answered from any meta-analysis, it is representative of the general variation of microarray study designs and laboratory protocols across research laboratories. In this case the choice of inclusion criteria is a reflection of the requirement to make a meta-analysis useful under pragmatically limited conditions, as well as an explicit biological focus on IFN-γ induced changes in murine macrophages irrespective of dose or time frame.

Detailed study descriptions are provided in chapter 2, table 3.1 provides a summary.

Table 3.1 Per-study group sample sizes

Study	Platform	Organism	Number of control samples	Number of IFN-γ samples	Treatment details	Publication
1	MGU74av2	Mouse (C57B/6)	3	3	10 u/ml for 24 hours	(Kropp, Robertson et al. 2011)
2	MGU74av2	Mouse (Balb/c)	3	2	100 u/ml for 6 hours or 48 hours	(Popkin, Watson et al. 2003)
3	MGU74av2	Mouse (Balb/c)	2	2	100 u/ml for 6 hours	(Popkin, Watson et al. 2003)
4	MGU74av2	Mouse (C57B/6)	2	1	10000 u/ml for 6 hours	(Thomas, Galligan et al. 2006)
5	Mouse430A_2	Mouse (C57B/6)	1	1	500 u/ml for 24-48 hours	(Helming, Bose et al. 2005)
6	Mouse430A_2	Mouse (C57B/6 and Balb/c)	7	7	50 u/ml overnight	(van Erp, Dach et al. 2006)

3.1.2 Data pre-processing

There are filtering and data processing steps that apply to some or all of the studies used as input for the three meta-analysis models. All data sets underwent the same basic quality control checks, processing and statistical analysis (see chapter 2, section 2.5) in order to ensure all data exist on equal scales and are of suitable quality. All six data sets were reduced to those 9812 gene probes that are in common to the different microarray chip platforms. Due to minimal sample size, two of the six studies are only eligible for the non-parametric Rank Product meta-analysis, reducing other methods to using data from four studies. Depending on the biological reliability of those two studies, the additional two studies may introduce bias towards or improve results from the RP method. However, with the emphasis on this thesis on the real-world application of meta-analysis, exclusion of two studies would support the evaluation of that objective less than it would support the comparison between statistical methods on a sample-for-sample basis.

3.2 Meta-analysis models

The meta-analysis models are chosen with a view to cover the combining of p-values, combining of ranks, and combining of effect sizes. These are Fisher's combined probability model (FP), Rank Product model (RP) and Effect Size model (ES), respectively. In addition to these main approaches, variants and potential improvements are also tested. In order to interpret results and visualisations, it is important to note that any microarray analysis differs from 'traditional' analyses in the sheer number of variables (often interchangeably referred to as genes, probes or features) to test, and conversely, the very limited numbers of observations (interchangeably referred to as samples or arrays) for each variable. For each gene, an independent meta-analysis is carried out. The implications of this are raised in the discussion section of this chapter.

Where samples sizes (*n*) are mentioned, in the studies used for meta-analysis here, these refer to the number of biologically independent samples (in this case, individual mice).

3.2.1 Fisher's combined probability model

Also referred to as Fisher's sum of logs method, this model (Fisher 1932) predates the introduction of the term "meta-analysis" by Gene Glass (Glass 1976). It relies on the idea that any number of independent statistical tests on the same research and (null) hypothesis can simply be combined, resulting in a new statistic that follows a chi-squared distribution with degrees of freedom equal to two times the number of studies. This model belongs to a family of models concerned with combining p-values from individual studies. Other representatives of this model family all combine p-values, but use normal or binomial distributions for p-values, add weighting or set minimum significance thresholds for individual studies (Stouffer 1949, Wilkinson 1951, Sidak 1967, Edgington 1972, Simes 1986, Zaykin, Zhivotovsky et al. 2002). Other models use the same principles, but combine per-study test statistics rather than their significance (Ghosh 2003), and

this comes very close to the effect size model described later in this chapter. It is possible that Fisher's method is more susceptible to bias from individual studies (publication bias, reporting of one-sided tests), but the other models listed above also make assumptions that in themselves cannot be confirmed, e.g. the presence of a large p-value in only one study representing bias rather than variation or experiment factors. Fisher's basic assumption of probabilities ranging from 0 to 1 holds true in this thesis, with probabilities calculated and known fully, i.e., not drawn from research papers or other external sources that likely do not report statistical significance above the 5% threshold. Fisher's is the most generic of the 3 models used here, the methodology allows for the combination of p-values from any source, e.g. it would be possible to combine a statistical test performed on protein abundance with a statistical test performed on gene transcription level, although this option is not explored here. An issue that is explored here is the dependence of Fisher's meta-analysis on the nature of the statistical hypothesis tests performed in the individual studies, e.g. parametric tests performed on non-normally distributed data would transfer a non-robust result into the meta-analysis. Given the specific studies used here, how do meta-analysis results compare when the individual studies have been analysed with a parametric or a non-parametric model?

For studies i…K and per-study result of statistical hypothesis testing p_i, Fisher's combination of probabilities is defined as:

$$\chi^2_{df} = -2\sum_{i=1}^{K} \ln(p_i) \qquad \text{(Eq. 3.1)}$$

Term *df* defines the degrees of freedom and equal 2*K*. In words, the negative scaling of a sum of log transformed probabilities by a factor of two results in a chi-squared distributed variable for which a p-value can be looked up in a chi-squared probabilities table with df=2*K*.

3.2.2 Rank-Product meta-analysis

Rank Product (RP) methodology was originally proposed as a robust method for identifying differentially expressed genes in individual studies (Breitling, Armengaud et al. 2004) but was later expanded to application as a meta-analysis methodology (Hong, Breitling et al. 2006, Hong and Breitling 2008). The rank product approach changes the focus from isolated gene-by-gene testing of hypotheses centred on differences is mean expression to a focus on biological expression fold-changes and taking into account the relative ranking of genes to one another. It has been shown (Yuen, Wurmbach et al. 2002) that the ranking of gene expression fold changes can be inherently more stable across multiple studies than fold changes themselves. RP has the simple aim of identifying genes that are consistently top-ranked (in their differential expression between experiment groups) across replicated samples from two or more biological conditions. This metric is improved on by assigning a statistical probability to the rank product statistic, which is based on comparing the observed rank product statistic to a permutation-based null distribution of rank product statistics. The methodology was originally intended to provide a level of robustness in the analysis of small individual studies, but was quickly and easily extended to work across a number of studies. Unlike Fisher's combination of probabilities, no hypothesis tests are applied in any of the individual studies, their contribution to the meta-analysis is the rank product estimate of each gene. And because there is no standardisation to the variance, it is possible (even if it may not be advisable) to include studies with just a single sample in a treatment or control group.

The notation for the implementation of this algorithm is different from that used by Breitling or Hong, for the simple reason that the authors provide sparse or changing notation in the original and related papers.

Following the removal of systematic measurement error from all individual studies through background correction and normalisation steps (see section 2.5 in chapter

2), there are *S* studies. Each study is a separate expression data matrix with *G* genes in rows and *T*+*C* samples in columns, where *T* and *C* denote the total number of IFN-γ and control samples in a single study, respectively.

The first step is to calculate, for each study, all possible pair-wise gene expression ratios (or differences on log scale) between (not within) control and IFN-γ samples. This ratio or quotient *Q* for gene *g* in study *i* for a pairing *k* of samples can be represented as the outer product between the two groups in each study (using the inverse of one group to achieve division).

$$Q_{gik} = \left(T_{gi} \otimes C_{gi}^{-1}\right)_k \quad \text{(Eq. 3.2)}$$

T_{gi} and C_{gi} denote, for a given gene, expression values from all IFN-γ samples in study *i* and all control samples in study *i*, respectively. The total number of sample-pairings between IFN-γ and control samples for a study is K_i, which is T_iC_i. After calculation of these expression ratios, the new data matrix for each study now contains *G* rows and K_i columns. As they can now be considered as a single data matrix for the purposes of a meta-analysis, the total number of expression ratios and therefore columns is $K_{tot} = \sum_{i=1}^{S} K_i$

Next, expression ratios are converted to ranks (largest ratio is 1),

$$r_{gik} = rank(Q_{gik}) \quad \text{(Eq. 3.3)}$$

followed by obtaining the rank product *RP* for each gene *g* across all sample-pairings *k* in all studies *i*. It should be noted that the study index is no longer required and subsumed by the index for sample-pairings across all samples (K_{tot}). The below rank product equation is also equivalent to the geometric mean of the ranked differential expression for each gene.

$$RP_g = \left(\prod_{k=1}^{K_{tot}} r_{gk} \right)^{1/K_{tot}} \quad \text{(Eq. 3.4)}$$

Subsequent to obtaining the observed RP statistic, a random experiment with identical numbers of samples and genes is created by randomly permuting all gene expression values within each array relative to gene identity. This is repeated *B* times, each permutation set followed by equations 3.2 to 3.4. Each permutation run yields a null rank product statistic:

$$RP_g^{*(b)} \quad \text{(Eq. 3.5)}$$

Over *B* permutations this constitutes the reference null distribution RP_g^* of *RP* values that are observed by chance. The exact statistical significance for a gene is the number of permutation estimates that are smaller than or equal to the observed estimate, divided by the total number of permutations:

$$p\left(RP_g\right) = \frac{\sum_{b=1}^{B} \mathrm{I}\left(RP_{bg}^* \leq RP_g \right)}{B} \quad \text{(Eq. 3.6)}$$

Finally, the procedure (equations 3.3 – 3.6) is repeated with rankings reversed, meaning there are separate runs for identifying up and down regulated genes.

Although the authors briefly discuss per-study weighting in the contribution to meta-analysis estimates by other models, this discussion on weighting is not extended to the RP method. Equation 3.4 corroborates that the RP model is implicitly weighting each study's contribution to the meta-analysis, because the

number of sample-pairings within it (K_i) grows with the square of study sample size, assuming sample sizes in each treatment group are equal (or more generalised, it grows with product *TC*). As each doubling of sample size increases a study's K_i and therefore the contribution to the meta-analysis rank product 4-fold, this is a strong implicit weighting factor. A modification of this strategy is therefore proposed and assessed here. A generalised weighting scheme based on study size can be expressed as follows, with rank values still going into an overall rank product, but weighted by study:

$$wRP_g = \left(\prod_{k=1}^{K_{tot}} r_{gk}^{w_k} \right)^{1/\sum_{k=1}^{K_{tot}} w_k} \quad \text{(Eq. 3.7)}$$

Term w_k is a vector of weights with length K_{tot}, with each element k in a study i weighted in relation to the number of sample pairings K_i in that study.

Two different weighting schemes are considered here, in addition to the standard method. If n is the per-group sample size of a study and study groups are balanced, then for the standard method individual study weight is n^2. In the first new weighting scheme, here referred to as proportional weighting, individual study weight is 2n. In the second weighting scheme, here referred to as equal weighting, each study's weight is 1. Formally, the two new weighting schemes are as follows:

Proportional weighting: $$w_k = \frac{T_i + C_i}{T_i \times C_i} = \frac{T_i + C_i}{K_i} \quad \text{(Eq. 3.8)}$$

Equal weighting: $$w_k = \frac{1}{T_i \times C_i} = \frac{1}{K_i} \quad \text{(Eq. 3.9)}$$

With the exception of sample size per group n=1 or n=2, the relationship between the normal weighting w_{norm}, equal weighting w_{equal} and proportional weighting w_{prop} is: $w_{norm} > w_{prop} > w_{equal}$.

Once the weighted rank products are calculated, comparison against a newly constructed null distribution of permutation-based weighted rank product values yields new probability values, as described for the non-weighted rank products.

With respect to computing p-values for rank product values (eq 3.5 and 3.6) through a permutation-generated null distribution, recent work on this subject (Eisinga, Breitling et al. 2013) suggests another method. The authors suggest that the above permutation strategy performs well for genes with large rank product values (resulting in less significant p-values), but is either less accurate in the generation of low p-values or only accurate for unfeasibly long computation times. The introduced method calculates the probability of a given gene's observed rank product by relating it to the total number of rank value combinations that could attain the same rank product value in a data set with a given number of genes and sample replicates. The exact p-value for this gene is then the probability of the observed rank product being equal to or less than the observed rank product.

3.2.3 Effect size model

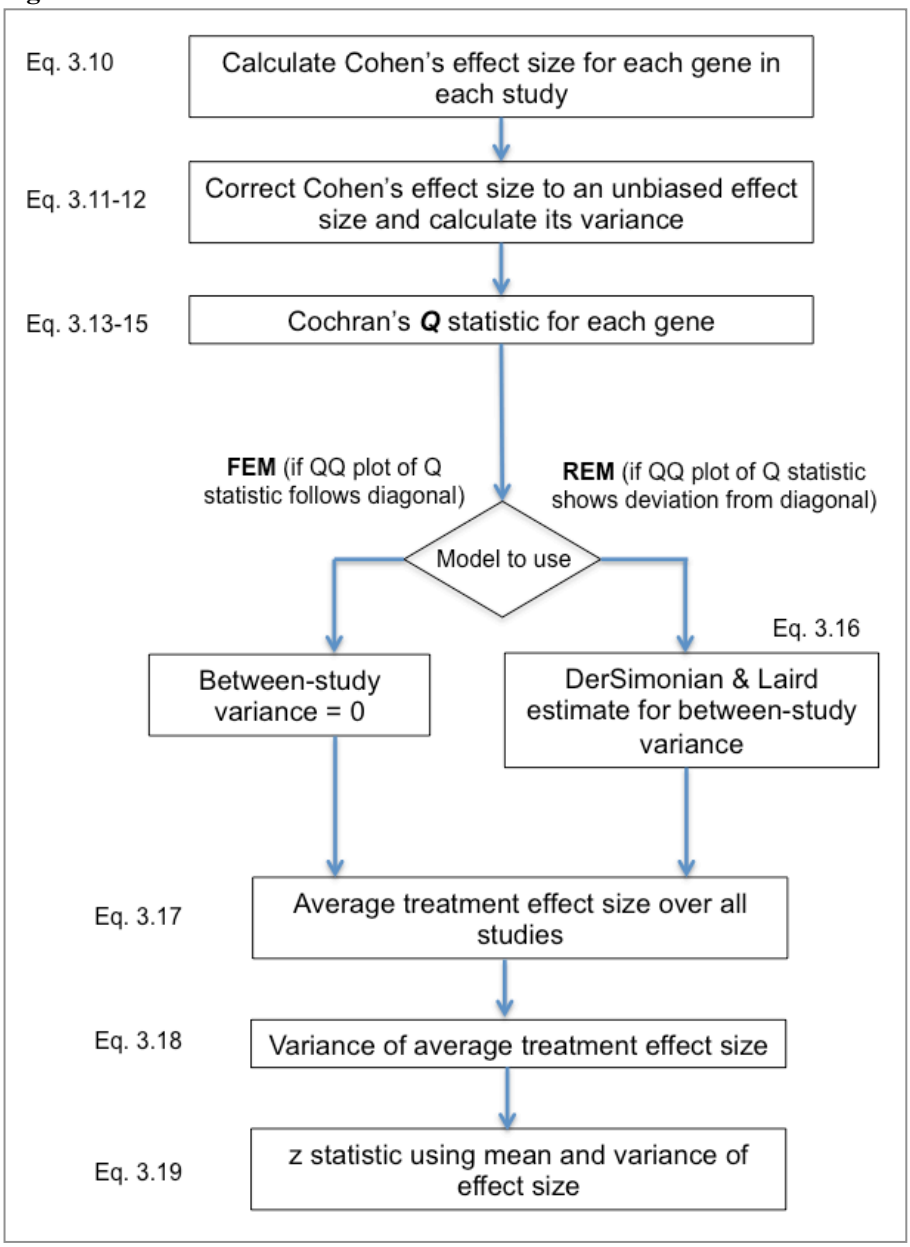
The effect size model (Choi, Yu et al. 2003) draws on a standardised difference in means between two biological conditions in a study in order to identify a meta-analysis effect size across all studies. It differs from Fisher's and Breitling's meta-analysis model mainly in its inherent use of within-study and between-study variability in the meta-analysis estimate, which is of course related to the size of a study. Effect size and its variance are calculated for each gene in each study, with the inverse of its variance used as a weight in combining the estimates for a gene across all studies, giving greater weight to larger studies or generally to studies with less random variation. Another feature of this methodology is the ability to test for between-study variance (Cochran 1954) and if required compute estimates for between-study variance (Dersimonian and Laird 1986). This allows a decision on the type of model that is used to combine effect sizes across studies, where the choice is between fixed effect model (FEM) and random effects model (REM).

For the REM, the assumption is that the studies in a meta-analysis are a random sample of a theoretically infinite number of studies with different true effect sizes (e.g. subject to multiple study population differences, known or unknown), all centring on an average true effect size. In this scenario, one does not assume all studies to be identical even in their theoretical outcome. For the studies in this thesis, one may for example not expect different mouse strains or interferon gamma doses to lead to exactly the same level of gene expression. As such, one would in the meta-analysis not only have to account for the variance within a study (depending on statistical sample size), but also for the variance between studies that is caused by the above-mentioned factors. When needing to determine an overall average effect size across all studies in a meta-analysis, under REM this average needs inverse weighting based on an estimated value for between-study variance (e.g. DerSimonian & Laird, as detailed in eq. 3.16) as well as the default weighting (addressing per-study random sampling error) by the inverse of within-study variance.

For the FEM, the assumption is that every study in the meta-analysis measures the same true effect size and all studies would therefore theoretically (with large sample sizes) return the same result. This scenario is really only plausible if all samples in all studies are drawn from the same population. If this assumption of a fixed effect were true for the studies used in this thesis, either only one mouse strain (or other identical experiment factor) would be considered, or one would assume that gene transcription levels are completely independent of the mouse strain (or other experiment factor) used. The between-study variance is then assumed to be 0 and the overall average effect size only requires inverse weighting by the within-study variance.

An overview of the full effect size (ES) model is provided in figure 3.1.2 below and the approach is detailed subsequently.

Figure 3.1.2 Effect Size model overview



This figure outlines the order of computations for the Effect Size meta-analysis model. "Eq." numbers refer to the equation numbers given in this chapter, FEM and REM refer to Fixed Effect Model and Random Effects Model, respectively. QQ refers to a quantile vs quantile plot, here specifically the quantiles of Cochrane's Q statistic vs the quantiles of the theoretical chi-squared distribution with N (number of studies) minus 1 degrees of freedom.

Among many possible options, the chosen estimator for effect size for a given study *i* is Cohen's *d* (Cohen 1988):

$$d_i = \frac{\bar{X}_i^T - \bar{X}_i^C}{S_i} \quad \text{(Eq. 3.10)}$$

Terms *T* and *C* refer to the treatment (IFN-γ) and control group, and S_i is the pooled standard deviation across both groups. Effect size *d* is obtained for each gene in each study. It should be noted that this estimator is similar to a t-statistic, but the difference in means is not standardised to the standard error but to the pooled standard deviation, thus sample size is not taken into account. Effect size *d* is then corrected (by scaling the biased effect size in relation to *N*) to an unbiased effect size estimate *d'* as introduced by Hedges and Olkin (Hedges 1985), where *N* is the total number of samples in a study *i*:

$$d'_i = d\left(1 - \frac{3}{4\left(n_i^T + n_i^C - 2\right) - 1}\right) \quad \text{(Eq. 3.11)}$$

A variance estimate for this unbiased effect size is obtained next, with n_T and n_C the number of samples in treatment and control groups, respectively:

$$\hat{\sigma}^2_{d'_i} = \left(\frac{1}{n_i^T} + \frac{1}{n_i^C}\right) + d'^2 \frac{1}{2\left(n_i^T + n_i^C\right)} \quad \text{(Eq. 3.12)}$$

Having established effect size and effect size variance for each gene in each study, it is now necessary to determine the type of statistical model required to combine estimates across studies, based on variation in study outcomes between studies. This heterogeneity can be estimated on the observed data, and although it has low statistical power with small numbers of studies in a meta-analysis (Gavaghan,

Moore et al. 2000), Cochran's Q statistic (Cochran 1954) is used for this purpose. For each gene, Q is defined as:

$$Q = \sum_{i=1}^{S} w_i (d'_i - \hat{\mu})^2 \qquad \text{(Eq. 3.13)}$$

Term S is the total number of studies. While d' has been obtained before this stage, per-study weight w_i and estimated treatment effect over all studies $\hat{\mu}$ are defined as:

$$w_i = \frac{1}{\hat{\sigma}^2_{d'_i}} \qquad \text{(Eq. 3.14)}$$

$$\hat{\mu} = \frac{\sum_{i=1}^{S} w_i d'_i}{\sum_{i=1}^{S} w_i} \qquad \text{(Eq. 3.15)}$$

Summarising the above steps, Cochran's Q is the squared difference between the estimated overall treatment effect and the estimated effect in individual studies, where individual studies are weighted by their inverse variance. The Q statistic follows a χ^2distribution with S-1 degrees of freedom, and while it is trivial for single-variable meta-analyses to obtain a p-value of statistical significance based on this distribution, for multi-variable microarray-based meta-analyses this process is adapted somewhat, by comparing the quantiles of empirically obtained Q statistics of all genes against the quantiles of a theoretical χ^2distribution with S-1 degrees of freedom. A quantile-quantile plot (QQ) is suitable to show if the empirical values are more extreme than expected, major deviations from the diagonal suggest that studies are heterogeneous and a fixed effect model (FEM) is inappropriate for combining effect sizes across studies, and a random effects model (REM) should be used instead.

If the statistical test result does not reject the assumption of study homogeneity, effect sizes are combined across studies under the FEM model, with between-study variance set to 0. However, if studies are shown to be heterogeneous, then effect size estimates across studies are combined under the REM model and an estimate for between-study variance is required. A weighted estimator for this purpose has been proposed by DerSimonian and Laird (Dersimonian and Laird 1986):

$$\hat{\tau}_{DL}^2 = max\left\{0, \frac{Q-(S-1)}{\sum_{i=1}^{S} w_i - \left(\sum_{i=1}^{S} w_i^2 \,/\, \sum_{i=1}^{S} w_i\right)}\right\} \qquad \text{(Eq. 3.16)}$$

Term *S* is the total number of studies and w_i is the inverse variance $1/\sigma^2$ for each study *i*.

With individual study effect sizes calculated, a decision on REM or FEM made through a QQ plot of Cochran's Q statistics, and an estimate for between-study variability of true effect size obtained, it is now possible to combine effect sizes for each gene across all studies. The combination provides estimates for the average treatment effect (in this case IFN-γ) and the variance of the average treatment effect. The average treatment effect size can now be calculated by weighting the per-study unbiased effect sizes with the inverse of the within-study variances and the estimated between-study variance (0 if FEM used):

$$\hat{\mu}(\tau^2) = \frac{\sum_{i=1}^{S}(s_i^2 + \hat{\tau}_{DL}^2)^{-1}\, d'_i}{\sum_{i=1}^{S}(s_i^2 + \hat{\tau}_{DL}^2)^{-1}} \qquad \text{(Eq. 3.17)}$$

The associated variance of the average effect size is:

$$Var\left(\hat{\mu}(\tau^2)\right) = \frac{1}{\sum_{i=1}^{S}(s_i^2 + \hat{\tau}_{DL}^2)^{-1}} \qquad \text{(Eq. 3.18)}$$

If FEM were the model under which effect size estimates were to be combined across studies, then between-study effect size variance $\hat{\tau}_{DL}^2$ would be 0 and therefore not have any influence on the outcome. For the four studies considered here, the QQ-plot (appendix A3) of observed versus expected Cochran's Q quantiles shows strong rejection of the null hypothesis of study homogeneity, requiring use of a REM model for combining effect size across studies, as described above. But in either case, standardising the average treatment effect size to the standard deviation of the effect size yields a z score as per-gene meta-analysis result:

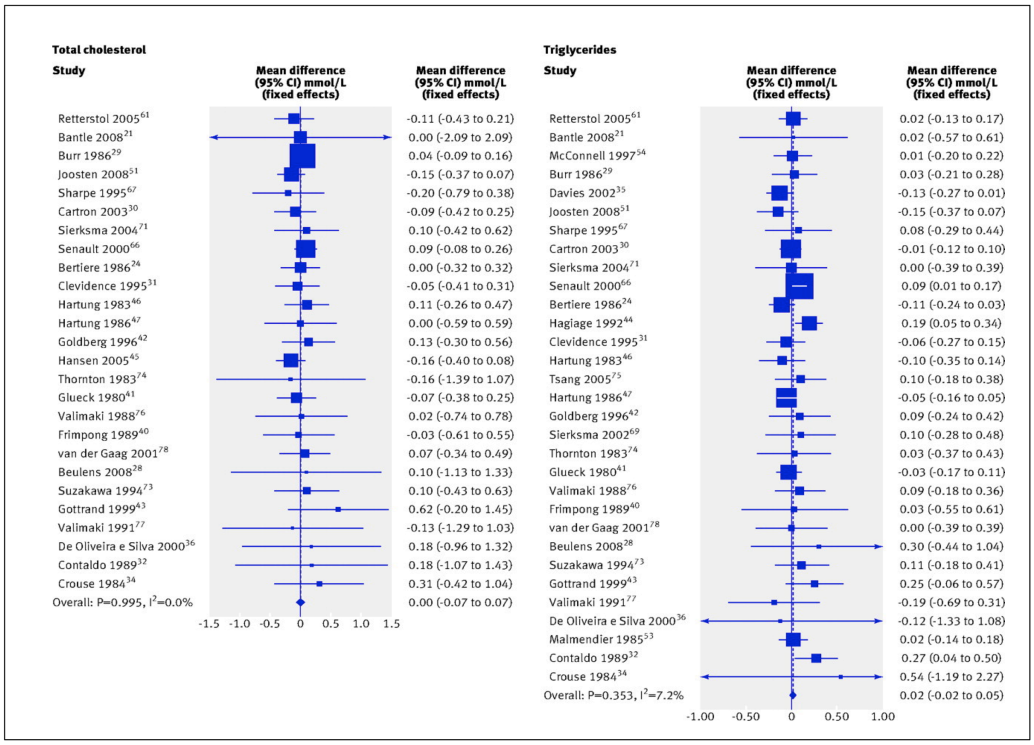
$$z = \frac{\hat{\mu}(\tau^2)}{\sqrt{Var\left(\hat{\mu}(\tau^2)\right)}} \qquad \text{(Eq. 3.19)}$$

Statistical significance values p can be derived from z by relating it to a standard normal distribution table, under the usual normality assumptions. Alternatively, a null distribution of z values can be constructed through permutation and repeat estimation of z on the original data, in cases where normality cannot be not shown or assumed.

3.3 Assessment methods for meta-analysis results

For meta-analyses performed on a single variable or very few variables, it is usually straightforward to assess the results in that one can simply list all relevant output (e.g. study contribution coefficients, statistical significance, confidence intervals) and plot results of individual studies and the meta-analysis in a standard forest plot as shown in the example, figure 3.2.

Figure 3.2 Example of a standard forest plot



(Brien, Ronksley et al. 2011). Brien et al, British Medical Journal 2011; 342. Fig 2; used under Creative Commons Attribution Non-commercial License.
The first column of each of these two panels lists the individual studies included in a meta-analysis, the second column graphically displays the analysis outcome of each study, using the same outcome measure (in this case, mean difference between two study groups) as that combined for a meta-analysis. The centres of squares indicate the mean difference between the treatment groups in a study, with horizontal lines indicating its 95% confidence limits. The area of the square is proportional to the study's weight in the meta-analysis. The diamond indicates the meta-analysis result across all individual studies, with the lateral edges of the diamond denoting the 95% confidence limits for this estimate.

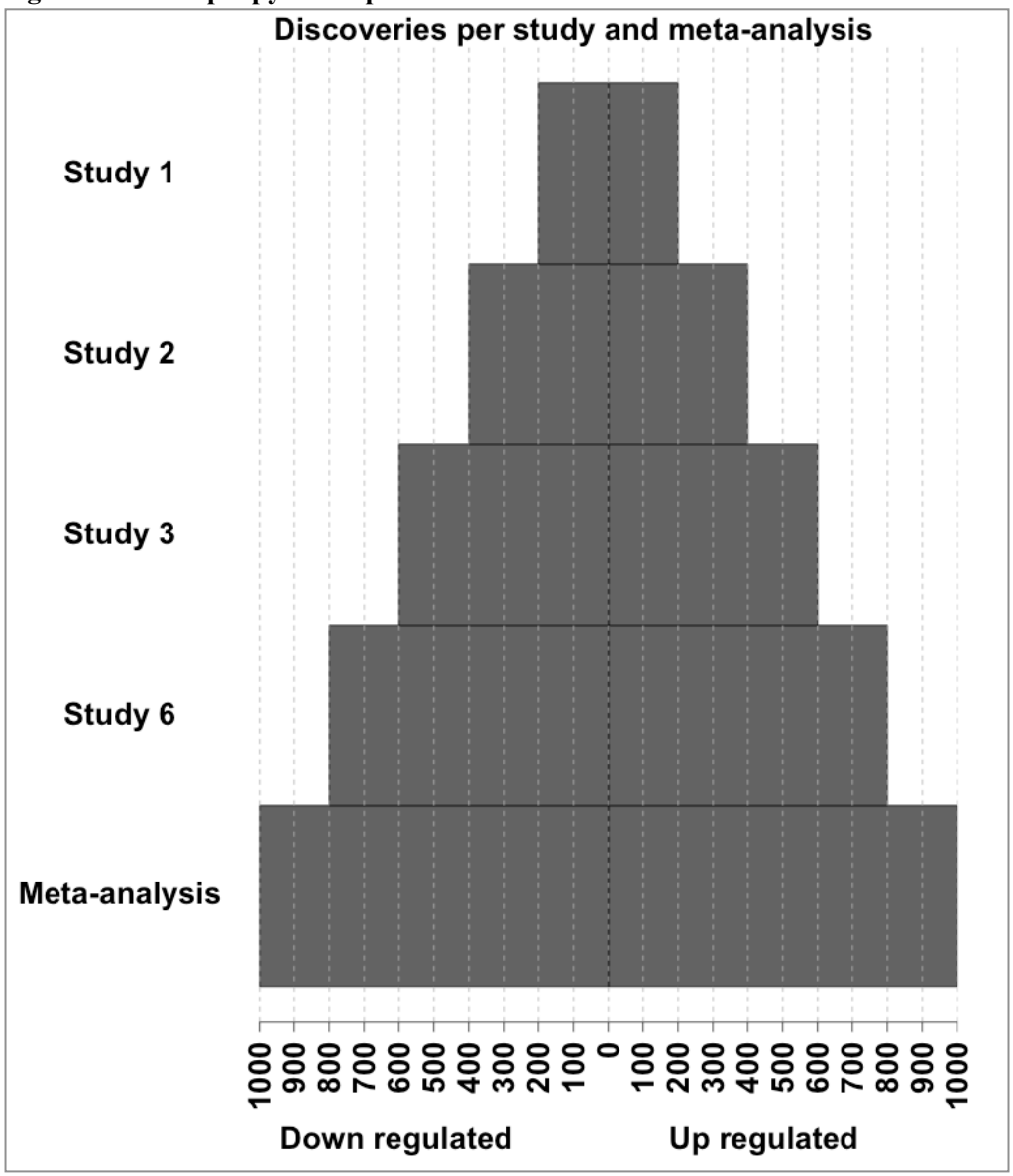
While this is a good representation, it does not accommodate microarray studies where thousands of variables (i.e. genes, probes, features) are analysed and one graph per gene would be required to represent this information. Analogous to heat maps representing microarray gene expression, relative expression or fold changes, it would be possible to represent per-study and overall effect sizes or equivalent statistics in a form of heat maps (or heat maps after hierarchical clustering on variables) where that effect size is mapped to a colour and shown for all genes across all individual studies and the meta-study. But also analogous to heat maps from basic microarray data, such an overview is difficult to assess quantitatively or meaningfully.

For this thesis, four principal methodologies have been used (and partially expanded on) to summarise these results. They provide numerical results but for simplicity and ease of categorisation are here named after their primary graphical output: Pyramid plots, IDD plots, ROC curves and CAT plots.

3.3.1 Pyramid plots and discovery counts

At the simplest level of result evaluation, note is taken of the number of "discoveries" made in individual studies and the number made through meta-analysis. The term discovery is at this stage not associated with any biological meaning, it is only intended to be a conceptual and short phrase to replace "number of statistically significant genes", where statistical significance is set at an alpha of 1%. Counts of discoveries are displayed in horizontal bar plots that distinguish between discoveries in up- or down regulation with reference to the control condition. As these bar plots come with some level of expectation that a meta-analysis identifies most individual-study discoveries as well as meta-analysis-only discoveries, the corresponding largest bar of counts would be at the bottom of the graph like the base of a pyramid. This type of graph is proposed within the framework of this thesis and not previously used elsewhere, a basic example is therefore shown in figure 3.3.

Figure 3.3 Example pyramid plot



Each horizontal bar represents the total count (indicated on X axis) of statistically significant genes in a given study or in the meta-analysis of the combined study results. The bar section to the left of vertical zero reflects down regulated genes (with respect to control condition), the bar section to the right reflects up regulated genes. If the studies are approximately sorted by number of samples (study size), the expectation would be that smaller study sizes (at the top) have less power to identify significant changes and will therefore be narrower. Similarly, a meta-analysis might be expected to increase the power to detect significant genes beyond that of any individual study and is therefore represented at the base of this 'pyramid'. Deviations from this pyramid structure may indicate issues or features of studies or analysis methods.

The visualisation of discoveries per study or through meta-analysis is informative about the relationship between study size and discoveries as well as the

relationship between individual studies and the meta-analysis in question, and it does address the need to summarise information across thousands of genes. However, simple counts do not provide any information on the magnitude of effect sizes or levels of significance, they do not map the outcome for individual genes in single-study analysis to the meta-analysis outcome for the same genes, and they do not allow a precise comparison between different meta-analysis methods. These points are addressed by the IDR method in the following section.

3.3.2 Integration-driven Discovery Rates (IDR)

A moderately more formal method of assessing meta-analysis benefits is derived from **I**ntegration-**D**riven **D**iscovery (IDD) and **I**ntegration-driven **D**iscovery **R**ates (IDR) as introduced by Choi (Choi, Yu et al. 2003) and with microarray meta-analyses in mind. As the name suggests, this is aimed at quantifying the benefit of meta-analysis by focusing on discoveries that do not occur in single studies but only in their combined analysis. The original method is adapted to work within the framework of this thesis, as the original was not intended for comparisons of different meta-analysis models. For this purpose, it is the negative log of significance p (one-sided) of the respective average treatment effect statistic that is plotted against the IDR, rather than a z score. This allows comparison between meta-analysis models on the same scale, with the negative log transformation aiding in visualisation of the otherwise compressed scale of p-values. Equations have been adjusted accordingly.

In detail, IDD is a count of the number of statistically significant genes (or "discoveries") in the meta-analysis that are not statistically significant in any of the individual studies, i.e. meta-analysis-only discoveries. The IDR is that count divided by the total number of discoveries in individual studies and the meta-analysis (total discoveries), i.e. the proportion of total discoveries accounted for by meta-analysis alone, with this in theory ranging from 0 to 1. For each individual gene, binary IDD status is defined as

$$IDD_g := \left[p_{meta} < p_{th} \; and \; \sum_{i=1}^{K} I(p_i < p_{th}) = 0 \right] \quad \text{(Eq. 3.20)}$$

Where p_{meta} is the statistical p-value resulting from the meta-analysis of gene g, p_i is the statistical p-value resulting from an analysis of study i, and p_{th} is a given threshold for statistical significance.

When this binary index is aggregated across all genes G as

$$IDD = \sum_{g=1}^{G} IDD_g \quad \text{(Eq. 3.21)}$$

, IDR is the number of meta-analysis-only discoveries in relation to the total number of discoveries:

$$IDR = \frac{IDD}{\sum_{g=1}^{G} [p_{meta} < p_{th} \; or \; \sum_{i=1}^{K} I(p_i < p_{th}) > 0]} \quad \text{(Eq. 3.22)}$$

Note that the contents of the Iverson bracket in equation 3.22 are not that from 3.20 but aggregate the total number of genes that are statistically significant in an individual analysis or the meta-analysis. The metrics obtained through this method are shown on IDR plots, which show the rate of meta-analysis-only discoveries (standardised to total number of discoveries as described above) for different thresholds of statistical significance.

IDR plots provide an immediate impression of the comparative performance of meta-analysis models, where a larger IDR indicates that the method is better in combining study results. However, the use of proportions can hide underlying absolute counts, and this graph therefore does not replace simple counts as shown in the pyramid plots. IDR may also be affected through the quality of analysis in

individual studies, e.g. if the Rank Product statistic is more robust at identifying statistically significant genes in small individual studies, the proportion of meta-analysis-only discoveries (IDR) is reduced. Both the pyramid plot and the IDR plot also rely on the assumption that every discovery is a good and biologically sound discovery. This assumption cannot be supported in light of false positive rates associated with simultaneous testing on multiple variables, and even a correction for this (by applying a Bonferroni or another p-value adjusting algorithm) does not assure biologically valid discoveries. This issue is addressed in the following section.

3.3.3 Receiver-Operating-Characteristics (ROC)

ROC is a generic technique for comparing observed to expected/known/true results, for example in evaluating new clinical diagnostics, evaluating algorithmic classifiers or evaluating algorithms applied to simulated data. By calculating how many outcome measurements of a new diagnostic or algorithm correctly match known true positive and known true negative outcomes, for any given level of that measurement, data for a curve combining true positive and false positive rates can be obtained. Calculating the area under this curve (AUC) provides a single summary estimate of how well the observed results match the expected results, with AUC=0.5 representing matching no better than random and AUC=1 representing a perfect method. In terms of assessing meta-analysis results, ROC is here used to compare observed discoveries to biologically expected results in form of reference lists of genes known to be transcribed when IFN-γ is present in murine bone marrow derived macrophages**.** The diagnostic input here is the vector of negative log transformed p-values (to avoid flipping of ROC curves) without distinguishing between up- and down-regulation. Any definition of "biologically expected" is likely to be far from comprehensive, because biological systems are complex and not readily reduced to a snapshot that would hold true for a given set of experiment conditions. This is also true in the specific case here, using "gold standard" reference lists consisting of accumulated snapshots of IFN-γ related genes. This is explained in detail in chapter 2, but briefly, there are two types of reference lists used here, one based on IFN-γ related genes retrieved from the

EntrezGene database, and one manually curated over several years at DPM. These lists will have omissions or inclusions depending on the sources they are derived from. As described in detail in chapter 2, in addition to the snapshot issue, another problems is that any such reference list by definition contains positive findings only, i.e. there is no definitive list of genes that are *not* involved in the IFN-γ pathway (other than simply assuming this is the whole genome minus reference list). Even within these limitations, a proportion of true biology captured and compared against would have advantages over no biology at all, at a minimum it is to be expected that any given analysis should recover at least a proportion of the already known underlying biology. In order to avoid circular reasoning in the biological interpretation of results, it is also important that reference lists be independent from the data used in the meta-analysis, and this is the case here. There are a number of disadvantages to the ROC evaluation. One is the uncertainty about the comprehensiveness of the gold standard reference lists, another is the assumption that a new method can only achieve performance as good as but not better than the gold standard. A third is that ROC in this case cannot distinguish between up-regulated genes (with respect to a control condition) and down regulated genes, because the gold standard lists do not contain this information. It is possible to compare up/down separately against the gold standard, but this would shift the emphasis on the regulation direction, not improve ROC itself. Lastly, while the threshold of declaring a discovery (statistically significant gene) is important in obtaining the continuous estimates for ROC curves, this provides no overview over the number of observed most significant genes compared to gold standard lists. The first three issues are problems inherent to the methodology, the last issue is addressed through the use of CAT plots.

3.3.4 Correspondence-at-Top (CAT)

This list-comparison statistic has been proposed for microarray-based meta-analyses (Irizarry, Warren et al. 2005). I visualises the proportion of meta-analysis discoveries that overlap with a reference list of known results. It differs from the other methods used here in a number of ways. Firstly, it has a focus on

"rediscovery" of known facts. As such, it is not limited to meta-analysis-only discoveries and counts meta-analysis discoveries irrespective of their significance status in individual studies. Secondly, the original authors present CAT plots in the context of what amounts to the leave-one-out principle (albeit without using more than a single permutation), where two studies undergo meta-analysis and a third study serves as reference. This raises questions, because it is basically a trade-off between an increased-power meta-analysis with less evaluation and a decreased-power meta-analysis with more evaluation. For this thesis, the method is adapted in that it does not compare meta-analysis results to a third study, but to the biological 'truth' through using the same biological reference lists as for the ROC analyses. The adaptation is subject to a small penalty in form of result granularity, because the biological reference lists are shorter than the full genome. Also, the displayed proportions are reduced to a selected threshold of statistical significance of $p \leq 0.01$, which can shorten plotted lines but provides useful comparative information on the number of discoveries per meta-analysis method.

3.4 Results

Results are outlined in terms of the assessment methods described in 3.3. Lists of statistically significant genes have not been shortened through correction for multiple testing, accepting the presence of false positives amongst results.

3.4.1 Discovery count results

In terms of simple counts of discoveries through meta-analysis (where they need not be unique to meta-analysis), there are several outcomes worth highlighting (figure 3.4).

Effect of study size. Statistical analysis of individual studies is strongly affected by study size (Figure 3.4, all panels), with the two smaller studies resulting in far fewer discoveries by any given statistical model. For the smaller studies, non-parametric methods (Figure 3.4, panels A-right, B) result in more discoveries than

parametric methods. Conversely, for the two larger studies parametric methods result in more discoveries than non-parametric methods (Figure 3.4, panels A-left, C). The prevalence of small microarray studies has long since led to many proposed analysis standards for small studies, favouring models that are either non-parametric (rank or list based) or moderate gene variance by borrowing information from other genes (Breitling, Armengaud et al. 2004, Smyth 2005, Jeanmougin, de Reynies et al. 2010). The small-study advantage of the non-parametric analysis may be due to the chosen one here being list based, while a simple rank based statistic (e.g. Wilcoxon Rank Sum test) would result in very granular p-values for such small sample sizes.

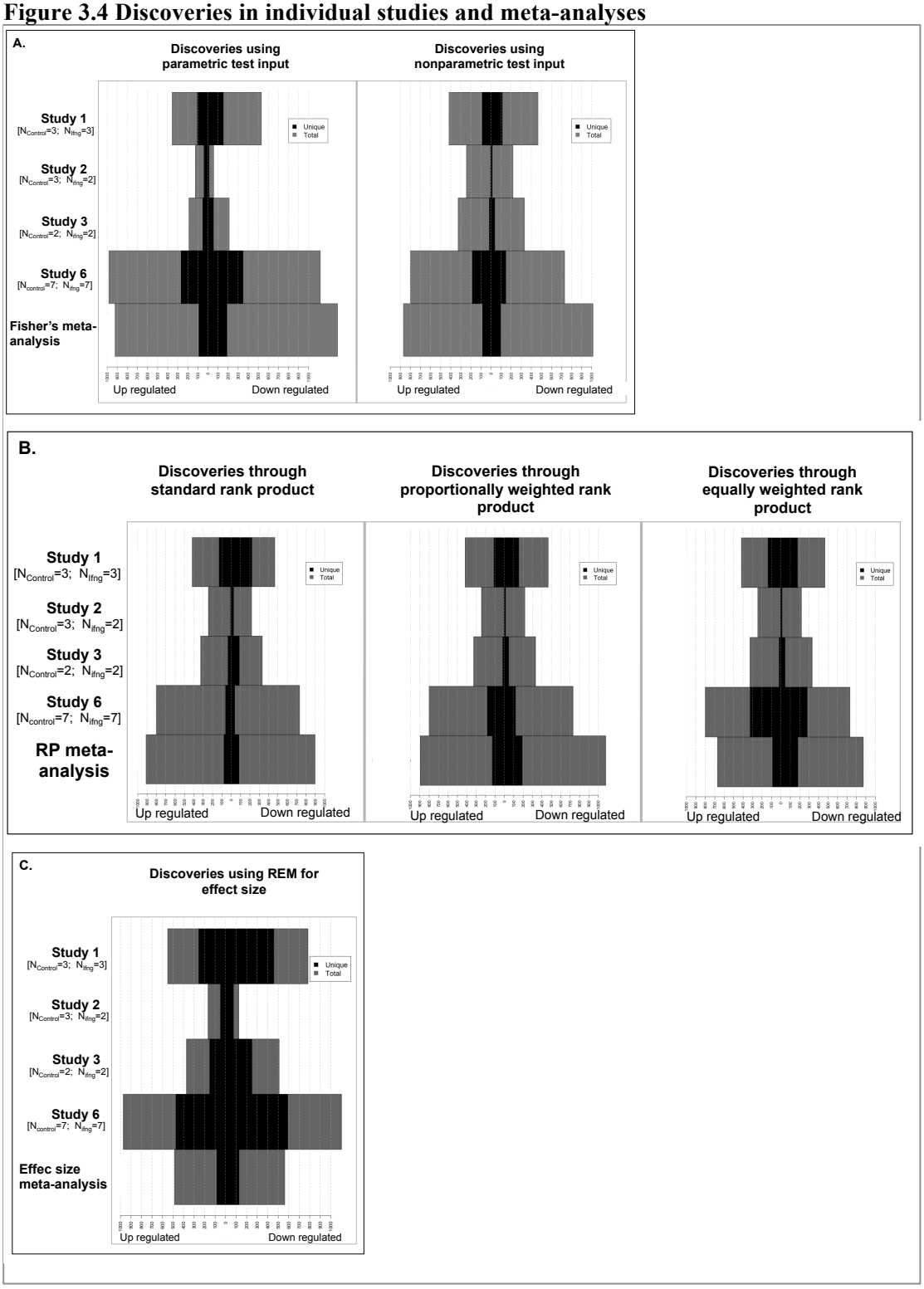
Effect size model. Comparing over all meta-analysis models, it is clear that the effect size model (ES) stands apart in terms of meta-analysis discoveries (it has far fewer) and in terms of discoveries unique to individual studies (far higher; black highlighted areas in figure 3.4 panel C). While this cannot be taken as a statement on the quality of these results, it highlights a potential weakness of the effect size metric in obtaining reproducible results across independent studies of differing size. This lowers the total number of meta-analysis discoveries including genes already identified in individual studies, but it does not greatly affect the number of meta-analysis-only discoveries (signified by the black regions in the meta-analysis counts), this subset is broadly comparable to other models.

Parametric vs. nonparametric generation of p-values for FP method. In line with above findings, when FP is used with parametric p-values, small studies result in fewer results than the nonparametric method, and vice versa. While the meta-analysis discoveries are almost identical (total count and unique count) for up regulation results, the non-parametric method results in fewer down regulation discoveries, with the largest study (for which the parametric model is better suited) likely driving those results, thus favouring the parametric model.

Effect of weighting on RP model. The model as proposed by the original authors is inherently weighted by the square of study group sample size (equation 3.4), this

will allow the largest study to have a very large effect on the meta-analysis outcome. While there is nothing to say that this is not a good weighting scheme (with larger studies expected to have lower variance), in case of unbalanced study sizes it outweighs the contribution of smaller studies and this is here considered against two alternative weighting schemes, one is proportional to study sample size and therefore moderates the effect of large studies more than the original, the other removes the influence of study size, with each study influencing the meta-analysis outcome equally. Up regulation discoveries are not affected by the proportional weighting, while down regulation discoveries are increased. Up regulation discoveries are strongly reduced by equal weighting, while down regulation discoveries remain similar to the original method. For both proportional and equal weighting, the number of discoveries made through meta-analysis only (i.e. not rediscoveries from individual studies) is higher for down regulated genes.

Figure 3.4 Discoveries in individual studies and meta-analyses



For each individual study analysis and the combined meta-analysis (Y-axis), the X-axis shows counts (out of total n_{genes}=9812) of genes with significantly higher expression in Ifn-γ than in controls (up regulated, shown left of vertical centre) and vice versa (down regulated, shown right of vertical centre). **X-axes in all panels are on the same scale, with plotting limits at count=1000 at either end from the central 0 count, and tick**

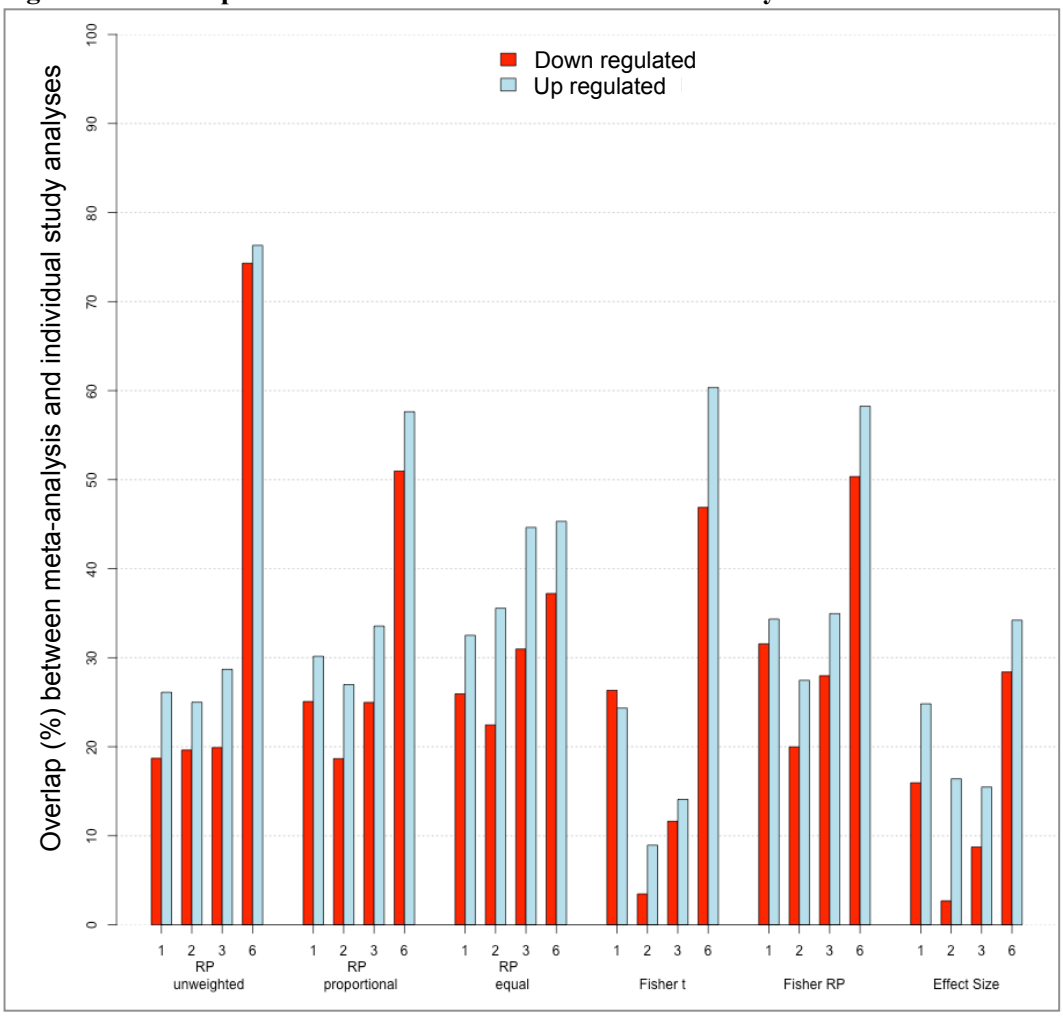
marks at intervals of 100. Black shaded areas within the horizontal bars indicate the number of statistically significant genes that were unique to that analysis.

Summary of quantitative assessment. Based on pure counts of discoveries, the suggestion at this stage is that, for the combination of small are large studies used here, the oldest and simplest model for combining study results (FP) appears to perform best, followed by the proportional-weighting version of the RP model. The worst performing meta-analysis method is the effect-size based model. The latter is not surprising given the small study sizes and statistical power issues associated with this. It is somewhat more of a surprise that the FP model based on parametric analysis results in the largest number of discoveries, ahead of the FP model with non-parametric study input and RP meta-analysis models, although the proportionally weighted RP method introduced here performs well. This is a reversal of findings by the original authors of RP (Hong and Breitling 2008) who do not recommend Fisher's method and observe RP to have the edge over both FP and an effect size model. Given that the same methods are used here (ignoring the new weighting schemes for now), this would suggest that meta-analysis results for small collections of small studies are very dependent on the composition, size and biological context of those studies. More studies are used here, and the single reasonably sized (in microarray study terms) study may be better suited to the simple parametric test and therefore drive meta-analysis results.

3.4.2 IDR results

In a level of complexity beyond simple per-method discovery counts, results uniquely obtainable through meta-analysis alone need to be identified. Before a detailed look at the meta-analysis proportion through IDR plots, a simple summary graph (figure 3.5) provides information on the difference between a meta-analysis and the analysis of an individual study in terms of number of discoveries. This differs from the results shown in section 3.4.1 (which do not identify overlap in gene sets) and the IDR plots in section 3.4.3 (which assesses results for changing values of p and reflects per-gene information across all individual study analyses).

Figure 3.5 Overlap between individual studies and meta-analysis results



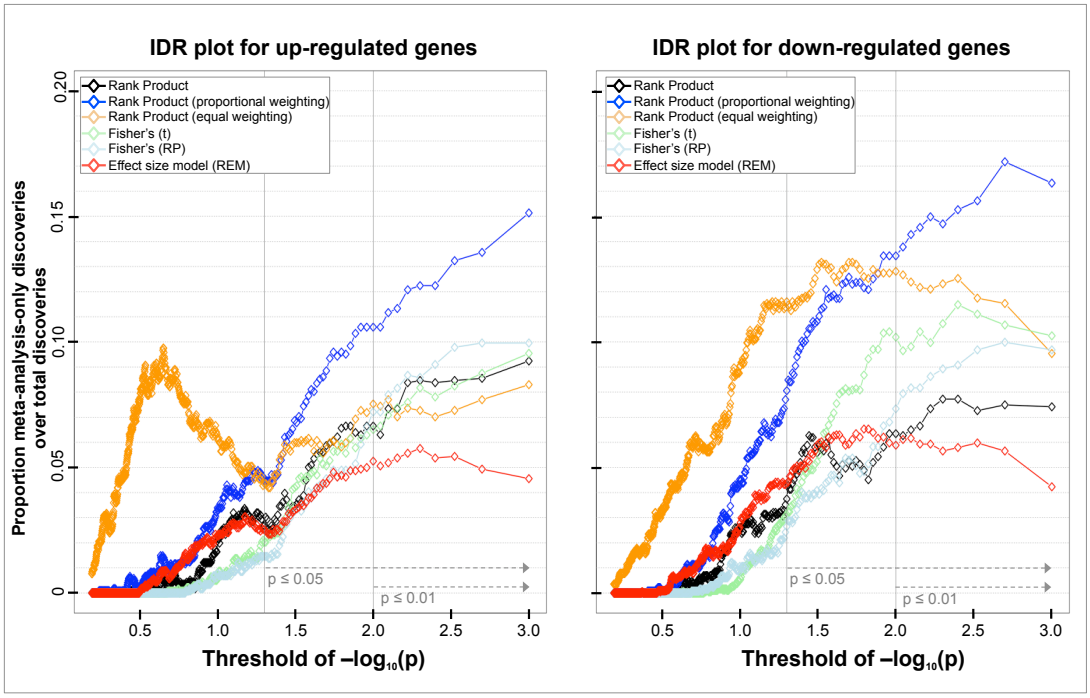
For each model of meta-analysis, overlapping proportion of discoveries (gene discoveries made at $p \leq 0.01$ in both single study and meta-analysis) is shown for the individual studies, analyzed by the method corresponding to that meta-analysis. The numbered columns refer to the study number. Red bars denote discoveries in down regulation, blue in up regulation with reference to control condition. A value of 100% is defined as the total number of statistically significant genes for a given meta-analysis model, and the shown percentages therefore reflect how many of these can be identified in a given single-study analysis.

Individual study performance compared to meta-analysis. From figure 3.5 it is clear that meta-analysis in every case is more than simply recall of single-study results, with even the largest study (#6) in the standard RP model identifying less than 80% of discoveries made in the corresponding meta-analysis. Non-parametric analyses all show comparable performance for the three smallest studies (#1-3),

while parametric methods have marked performance differences between the same studies.

The next step is to identify the number of genes identified in statistical meta-analysis that were not identified in any of the individual studies. Because the quantitative assessment in 3.4.1 indicates differences in discovery-numbers for up- and down-regulated genes, IDR plots are also broken down into those categories in figure 3.6 below. Crucially, this figure differs from figure 3.5 in that it aggregates information per gene across all individual studies rather than just stating how many genes are found to be statistically significant in a particular study.

Figure 3.6 IDR plots for meta-analysis-only discoveries



The X-axis represents a given threshold of statistical significance transformed to negative log scale for visualization purposes. The two vertical lines are visual guides at p≤0.05 and p≤0.01, respectively. Given a particular threshold on X, Y-axis shows the proportion of genes identified as significant in meta-analysis-only with respect to total number of discoveries (total = 1 = meta-analysis-only discoveries plus single-study discoveries in any of the individual studies).

Best model for up regulated gene set. For a commonly used statistical threshold of p≤ 0.05 (left vertical line), meta-analyses add from just over 1% to just under 5%

to the number of genes already identified as discoveries in individual studies, with only the two new weighting schemes for RP at the top end of this estimate, and the FP model using non-parametric tests as the worst performer. It must be noted that this could also mean the analysis of the individual studies is far better than that of other methods, this is outlined in the discussion below. At a more stringent statistical threshold of $p \leq 0.01$ and consequently including fewer false positive discoveries, it is clear that the proportional weighting version of the RP meta-analysis performs much better (IDR~10.5%) than the original RP model or any other model, with ES coming in last (IDR~5%). This pattern gets more pronounced if the statistical stringency for significant differential expression is increased further, as would be the case when adjusting p-values for multiple testing. The noticeable peak of increased proportions shown for the equal-weighting version of RP is in the region of not usually used p-value thresholds, meaning a discovery as defined here would be exceedingly easy to make and contain a large proportion of false positive results. This does not explain why it is different from the other methods in this abnormality, but this could be related to the weighting scheme increasing the weight of smaller ('bad') studies and decreasing the weight of larger ('better') studies, leading to a larger number of missed (at least at such a low threshold of statistical significance of $p \leq \sim 0.2$), discoveries in individual studies and therefore driving up the rate of meta-analysis-only discoveries. Given that the IDR drops again, this would point to the presence of a set of genes with somewhat variable expression properties in at least one of the studies.

Best model for down regulated gene set. Within the constraints set by microarray technology, IFN-γ appears to be associated with a larger number of down regulated than up regulated genes, which is seen in the generally higher proportion of meta-analysis-only discoveries. The pattern also seems to suggest that it is only in the set of down regulated genes that differences between meta-analysis models are much more distinct even for lower thresholds of statistical significance. At the modest statistical threshold of 5%, it is the equally weighted (IDR~11.5%) and proportionally weighted (IDR~8%) RP versions that show the largest meta-

analysis benefit, with all other models below an IDR of 5% and the non-parametric version of FP showing the least benefit. At the more stringent statistical 1% threshold, the proportionally weighted RP provides an IDR of ~13.5% and then continues to outperform other models at all higher statistical stringencies. Comparatively, the ES model provides the least benefit at higher statistical stringency.

Individual study results not rediscovered through meta-analysis. The reverse scenario of genes identified in any individual study but not through meta-analysis is also briefly investigated. They broadly answer the question, which of the 4 studies produce results that are not reproduced in a given meta-analysis. The corresponding graphs (appendices A6-10) are similar in construction, in that they show the proportion of genes identified in a single study that were not identified in the meta-analysis, across thresholds for p (in this case not transformed). Although there are many features in these graphs (e.g. differences in up and down regulated gene sets), the one that may best summarise this outcome is the comparatively high proportion of genes identified in the largest study (study #6) using the ES and FP models, particularly up regulated genes (blue). This would suggest that these meta-analyses are unable to re-discover a larger number of genes that were successfully identified in the largest study, presumably because the analyses of the three smaller studies outweigh the contribution of the largest study to the overall effect size and combined probabilities. Conversely, the RP meta-analyses – particularly the proportionally weighted and standard method – all appear to recapture a larger proportion of genes that were identified as statistically significant in any single study. A conclusion of this may be that the robustness of the RP methodology in a scenario with multiple smaller studies outweighs the parametric advantage of a single larger study contributing to the combined effect sizes or probabilities. In other words, where RP is concerned, more small studies may be better for the meta-analysis than single large studies. Where ES and FP are concerned, the meta-analysis outcome is negatively influenced by lower-powered studies.

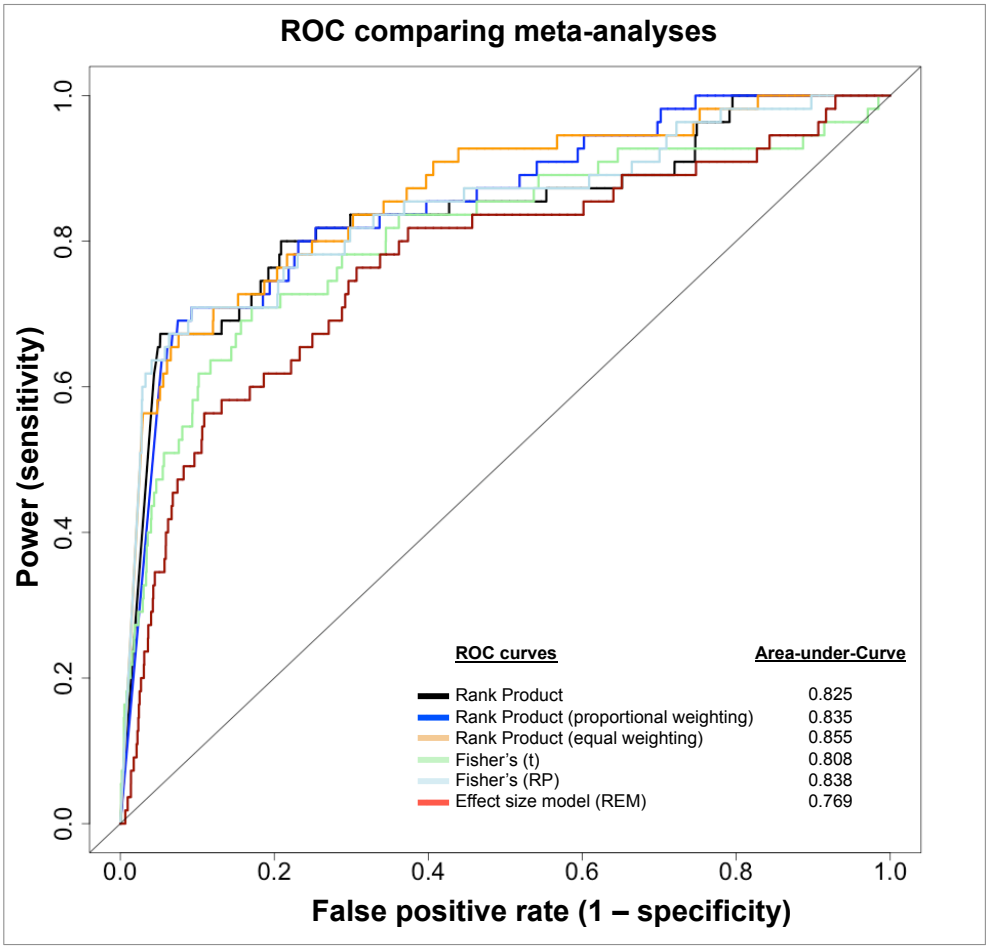
3.4.3 ROC results

Following the quantitative benefits of meta-analysis as described by discovery counts and IDR, it is now necessary to begin comparing results with reference to known biology, at least in broad terms. Detailed biological results are discussed in chapter 5. As an aside, comparison of a manually curated IFN-γ reference list and computer/database derived reference lists clearly shows that – although better than random with AUC > 0.5 – the latter do not contain the same quality of information as the former (appendices A4 and A5). This either signifies that the information contained in all-purpose-databases is far too weak to be of any use as a reference for any specialised experiment, or maybe more likely it signifies that a much more detailed database query structure than attempted here is required to produce a valid reference set for a specialised experiment, enriched for only those genes that are related to the specific biological hypothesis tested in the selected microarray studies. This description in practice matches the manually curated DPM gene list, motivating its use instead of the inferior lists generated through computational searches of the NCBI database.
Given the meticulous and deliberate construction of this list (section 2.7.2 in chapter 2) by several Division of Pathway Medicine MSc students over several years, this list is here considered as the "gold standard" to compare any new findings to, although this interpretation may be subjective to DPM. ROC results for all meta-analyses are directly compared in figure 3.7.

With all possible values of the negative log transformed statistical significance p sequentially serving as a threshold for determining a meta-analysis discovery, sensitivity and specificity are calculated with respect to the 'known' truth of discoveries in form of the DPM reference list of genes known to be involved in IFN-γ pathway in murine macrophages. All pairs of specificity (or rather 1-specificity, which is the false positive rate) and sensitivity (which is the true positive rate) are plotted, resulting in the ROC curves. The area under the curve is calculated as a representative outcome statistic of the curve, where a value of 0.5 represents the diagonal (sensitivity and false positive rate are equal for all possible significance thresholds, i.e. the overlap between meta-analysis results and

reference list is random) and a value of 1 represents a curve apex at the top-left coordinate (full sensitivity and specificity for all possible significance thresholds).

Figure 3.7 ROC curves for meta-analysis compared to DPM reference



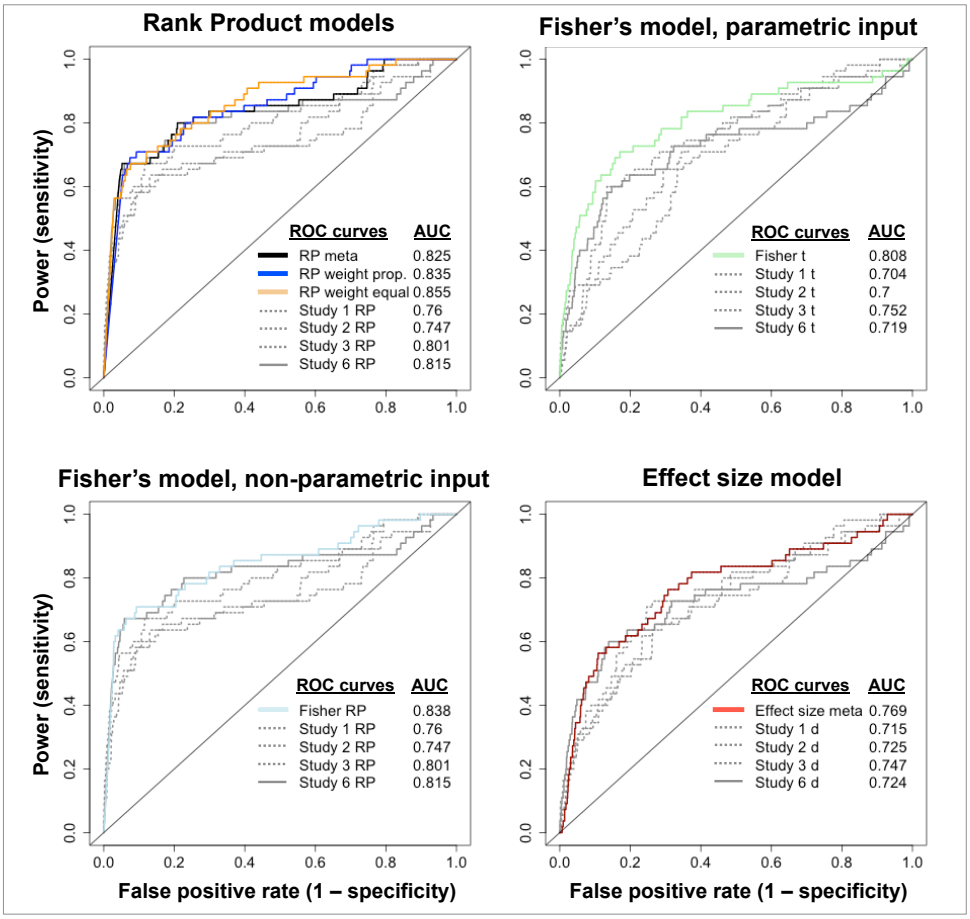
The curves in this ROC plot show sensitivity and false positive rate for all possible meta-analysis significance levels for which the meta-analysis outcome (significant or not significant) is compared to 'known' truth (contained in DPM reference list or not contained in it). For each meta-analysis method, the area under the curve is also shown as a summary statistic, where a value of 1 would indicate 100% sensitivity and 0% false positive rate for any given decision threshold of statistical significance

Meta-analysis performance comparison based on ROC. From this analysis it is clear that all meta-analysis findings broadly recover the known IFN-γ pathway. Perfect classification (100% sensitive, 100% specific) can of course not be expected from any method, given the gold standard list uncertainties described earlier. It is also clear from this analysis, that if the aim is to keep the rate of false positive findings low at below 10-20%, it would be more appropriate to choose

one of the RP meta-analysis models or the non-parametric testing for the FP model, as they are approaching around 80% sensitivity at that level. The sensitivity of the ES and the parametric FP model are 10-20 percent points lower for the same false positive levels. Apart from the comprehensiveness of the DPM gold standard list, another issue to be kept in mind in this ROC interpretation is the definition of false positive results, this is detailed in the discussion section of this chapter.

Individual study performance based on ROC. Since all statistical outcomes can be obtained for each study, it is also possible to assess their biological context through ROC curves, providing some insight as to how much value a meta-analysis adds over individual studies, particularly larger ones. Figure 3.8 repeats the meta-analysis ROC curves, but also includes for each approach the ROC outcome for individual studies.

Figure 3.8 ROC curves for study results compared to DPM reference



The curves in this ROC plot show sensitivity and false positive rate for individual study results and meta-analysis results. The analysis outcome (significant or not significant) is compared to 'known' truth (contained in DPM reference list or not contained in it). For each analysis method, the area under the curve is also shown as a summary statistic, where a value of 1 would indicate 100% sensitivity and 0% false positive rate for any given decision threshold of statistical significance. Grey lines (dotted and solid) refer to analysis results for single studies, the largest of these is a solid line, and the smaller studies intentionally all share the same dotted line pattern in order to increase visual contrast to meta-analysis results.

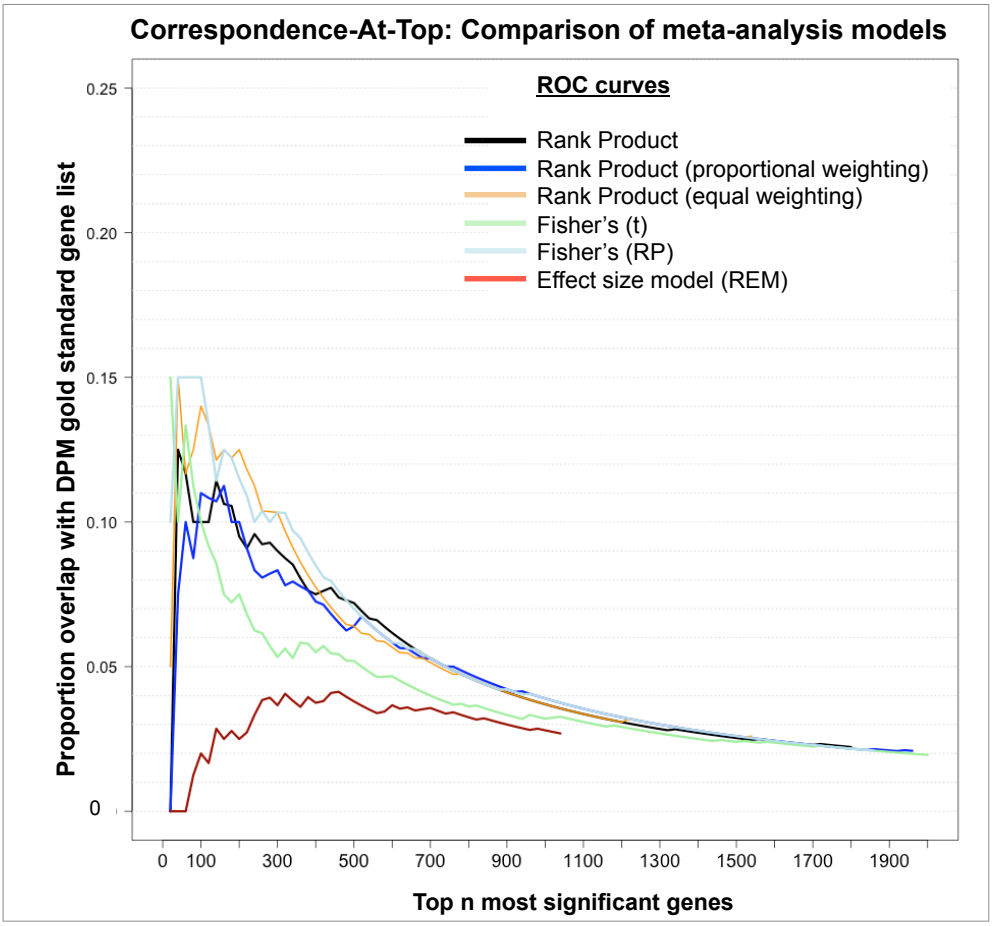
This assessment indicates that in terms of AUC the largest study (study #6) is in every case marginally worse than the corresponding meta-analysis, apart from the parametric FP where it is considerably worse. This is also true when limiting results to lower false positive rates only. On the whole this would therefore suggest that meta-analysis provides a marginal improvement over the statistical analysis of a single larger study, leading to two possible conclusions. One, meta-analysis may be superfluous in scenarios where a large study on the chosen subject exists. However, this would be outweighed by a second conclusion: Any value of

meta-analysis in scenarios where larger studies are indeed available may not lie in the identification of large numbers of previously unknown contributing genes, but only in the identification of a small number of previously unknown contributing genes. This is not entirely unexpected, because it is unlikely that after decades of research into IFN-γ there would be any reason to expect the discovery of a large number of new genes in the biological pathways identified to date. And it can certainly be assumed that genes with low but consistent expression changes are of low incidence but high value.

3.4.4 CAT results

Correspondence-At-Top assessment as used in this thesis is somewhat similar to ROC in that it uses the same biological reference list, but it highlights the number of ordered significant results that overlap with the reference list. Figure 3.9 compares this metric between the meta-analysis models.

Figure 3.9 Correspondence-At-Top using DPM reference



For a given number of the top n most significant genes (ordered) in a meta-analysis (X-axis), the proportion of genes in that set overlapping with the DPM reference gene list is shown (Y-axis). Each list is curtailed at significance level of $p \leq 0.01$, determining the length of a line.

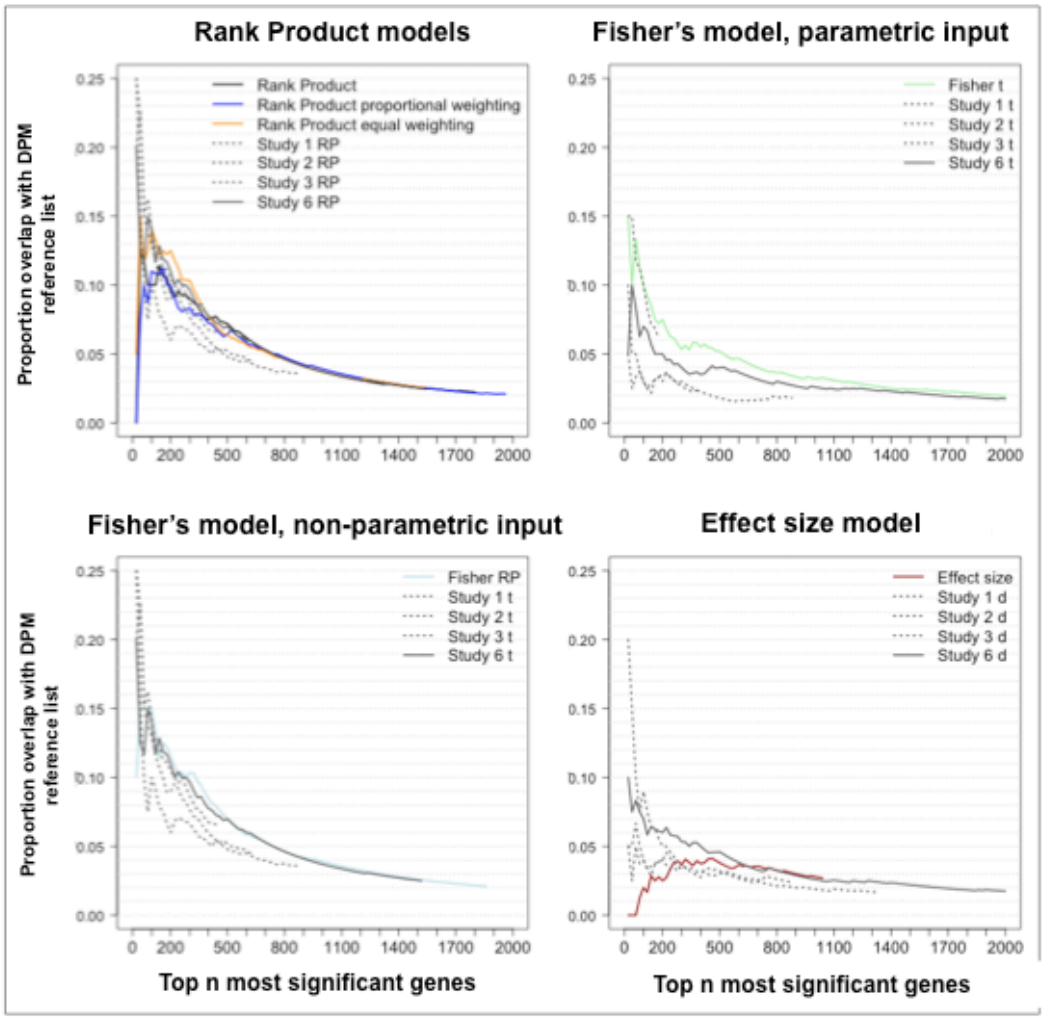
Meta-analyses reflect biological knowledge. Curves peaking towards the left indicate that the most significant genes are also the ones that have the best overlap with the DPM reference list. This is a welcome confirmation that meta-analyses in general reflect biological knowledge, although the ES model may be an exception to this, at least for the given of studies.

Gene set sizes. Optimum number (peaks) of top significant results is between 50 and 150 for most models. Beyond this, overlap with "known truth" decreases. An exception to this is the ES model, where this peaks at just under 500 or in plain terms, it is worse than other methods but for that reason does not deteriorate further.

Meta-analysis model performance. An unknown point from section 3.4.1 was the association of the ES model with the lowest number of meta-analysis discoveries, as that low number may still have had the best overlap with biological knowledge. Based on the CAT assessment for the ES model, this is not the case and the low number of statistically significant meta-analysis genes is also the worst reflection of known biology. It only approaches the biological validation level of other models once their result lists are extended to the top 1000 genes and have therefore accumulated more false positives. While both variations of the FP model show good performance where the top 100 most significant genes are considered, the parametric version quickly accumulates false positive results and is reduced to ES level performance. As for best performance with relation to biology, the non-parametric FP model and the equal weighing RP model are associated with the highest overlap rates and slowest inclusion of false positive results, followed by the proportionally weighted and standard RP models.

Individual study performance based on CAT. Similar to the ROC analysis, this method also allows a closer look at the performance of individual studies' analyses in relation to meta-analysis (Figure 3.10).

Figure 3.10 CAT for per-study results using DPM reference



For a given number of the top n most significant genes (ordered) in a meta-analysis (X-axis), the proportion of genes in that set overlapping with the DPM reference gene list is shown (Y-axis). Each list is curtailed at significance level of p≤ 0.01, determining the length of a line.

Based on this, the top ~100 most significant genes identified in individual studies (including smaller ones) appear to be a better match to existing biological knowledge than their corresponding meta-analyses. A more extreme case is the ES model, where all individual studies perform better than the meta-analysis when considering up to the top 300 genes. For all models, it is worth pointing out that it is not the largest study that best matches known biology, suggesting that some studies may have a more 'typical' study design in terms of interferon doses, time points or other protocols better matching the known interferon/macrophage biology as curated by DPM staff and students. This advantage of individual

studies does not extend to the inclusion of larger numbers of significant genes, it is here that their false positive count increases and therefore reduces overlap with the biological reference list. Meta-analyses are of course intended to increase the statistical power of analysis, and the reduced rate of false positive results among the top few hundred genes (beyond the initial top 100 as described above) when compared to individual studies seems to bear this out. As with ROC analysis, for the largest study the overlap with known biology is very close to that of the meta-analysis (RP models and nonparametric FP model), highlighting the need to identify a small set of genes for which the meta-analysis gain in statistical power is sufficient to identify genes that are not individually significant in even the largest study.

3.5 Discussion

Wider implications of these quantitative results and the biological investigation from chapter 5 and 6 are discussed in the final thesis chapter.

As outlined in chapter 1, the basic models of meta-analysis used here are not novel in themselves and the comparison between them is no longer novel since the publication of a paper (Hong and Breitling 2008) based on the same foundation works (Rhodes, Barrette et al. 2002, Choi, Yu et al. 2003, Breitling, Armengaud et al. 2004) as this thesis, with the added advantage that one of the meta-analysis models was their own. There is therefore a definitive need to highlight the areas in which this thesis differs, these being a) notional improvements to the actual meta-analysis models, b) inclusion of more studies (other papers use effectively two small studies, with a third used only as reference), c) a different biological system, c) result and validation against biological domain knowledge rather than simulation or reference studies, and d) an investigation of missing values and assessment of methods dealing with them.

Main outcomes. From discrepancies between simple counts of meta-analysis discoveries and evaluations taking into account gene identity or biological context,

it has to be concluded that the former are not a good indicator of performance. For example, the parametric FP model is associated with the highest number of meta-analysis discoveries (figure 3.4), but clearly also results in a larger proportion (better only than the ES model) of false positive results than other models (figures 3.7 and 3.9). It seems sensible to attach more weight to carefully curated information (DPM reference list) than to blind counts, the majority of interpretation is therefore based on results beyond counts.

From some results (figure 3.4) it would appear that a single study is sufficient to make the majority of discoveries that a meta-analysis is capable of. However, additional findings from CAT analysis show this to be only evident for the most highly significant genes (figure 3.10). The more the significance criterion is relaxed the more meta-analysis outperforms individual studies (while still staying within common statistical significance thresholds). The other issue of course is that even if a single study almost matches the outcome of a meta-analysis, the biological value lies in identification of the small number of added genes, with large sets of new biological findings not expected. This is particularly clear from figure 3.5, which identifies individual study's overlap with meta-analysis results and measures this at less than 80% at best.

Based on all available evidence so far, it is the RP models and the simple FP model (non-parametric variant) that perform best in a direct comparison of meta-analysis models. With reference to biology, the improvement over the largest single study alone is marginal (figure 3.8), but potentially still translates into dozens of new discoveries among hundreds of known ones. By other indicators (figure 3.10) and if using more stringent statistical thresholds (i.e. allowing fewer false positives), RP meta-analysis, or rather the two introduced weighting variants for RP, is more clearly adding value to identifying 'novel' genes. It may be the discrepancy between these two outcomes that in the end identifies new biology, because those novel findings or a percentage thereof would clearly not yet be present in the gold standard list and therefore not yet taken into account in the above marginal findings with reference to known biology. In the presence of outlier-sized studies in the meta-analysis, the alternative proportional weighting scheme for the RP model is of particular interest, as it by some measures (figure

3.6) shows promise, although with reference to biology this is more marginal and only when allowing for small false positive rates (figure 3.7). The better non-parametric meta-analysis outcomes can conceivably be attributed to their inherent robustness where the study sample size is low and unbalanced, variance high, gene expression not necessarily normally or log-normally distributed and the number of studies small and heterogeneous. For the RP model only, a small factor in the better results may be the inclusion of two more studies, although the contribution in terms of estimates is small, as explained later in this section. However, where evidence supports this for this set of circumstances and the particular models used, this should not automatically be generalised to scenarios where there is a larger number of small studies, a small number of large studies, a more homogenous mix of studies, or a more confined range of study inclusion criteria. Similarly, other non-parametric tests may be worse and other parametric tests may be better than those used here, with particular emphasis on those using information from multiple genes to moderate a parametric statistic, as they have shown their worth in general microarray analysis, e.g. empirical Bayes moderated t statistic (Smyth 2004).

The above main findings exist in context of a large number of issues that can affect them, these are discussed in the following paragraphs.

Single-study analysis methods. When interpreting single-study findings against meta-analysis findings, the comparisons are here made like for like, i.e. RP meta-analysis is compared to RP study analysis. Although there are widely used statistical tests – particularly the empirical Bayes moderated t test - for individual microarray studies, there is no single 'best' test when considering the different sample sizes here. In order to avoid comparisons of all individual tests methods to all meta-analysis models, like for like as explained above seems most appropriate. This may lead to some single-study vs. meta-analysis comparisons being subject to low-quality single-study tests, although this is mainly an issue for the effect size model, where the effect size statistic per study is inferior even to a Student's t test in that it does not pool gene standard error between conditions.

Alternative meta-analysis models. The models applied here are a conscious selection based on articles published early in the history of microarray meta-analyses. In the years since, the interest in microarray meta-analysis as a subject has increased, peaked, and decreased again, at least judging by annual citation counts with relevant terms (see chapter 1, figure 1.3). Despite this, a recent comprehensive review (Tseng, Ghosh et al. 2012) still refers to the three methods chosen here as main comparators. Other methods are being developed and may show benefits, e.g. Bayesian statistic-moderating techniques using multiple genes to model priors (Marot and Mayer 2009), but show no clear dominance of use.

Data integration techniques. Apart from meta-analyses in the statistical sense, a number of interesting methodologies have been devised that could also be used to obtain combined results across studies (outlined in chapter one). These often focus on the stability of biological motifs or the consistent co-expression of genes across multiple studies, some allow the integration of data from multiple domains, e.g. transcriptomics and proteomics. These are worthwhile pursuing in the same way that pattern analysis for microarray data is worthwhile pursuing in conjunction with statistical hypothesis testing, with the main difference being the same: Single-gene hypothesis in contrast to gene-gene relationships.

Study matching and gene identity. By limiting studies to Affymetrix-branded microarrays and using an Affymetrix algorithm that uses actual sequence similarity to combine data across their proprietary chip platforms, the problem of how to match genes between studies is avoided for this thesis, and the assumption can be made that the same gene is measured in all studies used here. However, this is of more concern if there is a heterogeneous mix of microarray platforms. The two problems posed are that a) it is unclear if probes meant to represent the same gene on different arrays are really measuring the same biological entity and b) that different measuring scales may need to be combined across studies if one is based on absolute levels of transcription and one on levels of transcription in relation to some control condition. This problem used to pose greater practical difficulties

and required more time solving in years past, but recently R/Bioconductor packages like MADGene have become available that automate this procedure somewhat (Baron, Bihouee et al. 2011).

Definition of false positives when using biological knowledge. When interpreting the overlap of meta-analysis results with a biological reference list, false positive results are related to two factors. One is the level of correctness and comprehensiveness of the biological gold standard list(s). This is addressed at length in chapter 2. The other factor is the meta-analysis itself, which aims to identify genes with small expression level differences on top of easier to identify genes with large expression levels. Genes with stronger transcriptional response will have a better chance of being identified in individual and smaller studies and can therefore be expected (with higher likelihood, at least) to contain true and known positives. Genes with weak but consistent transcriptional response are more likely to be unknown and not contained in the reference lists and therefore incorrectly counted as false positive results. This has to be considered in the interpretation, but the problem is of less importance if accepting that a) the number of genes that is uniquely identified by meta-analysis is small compared to what is already known and b) that some of these meta-analysis-only discoveries will already have been identified in studies outside this meta-analysis and therefore are part of the reference list(s).

Experiment factors. The discussion of the influence of study group sample size on meta-analysis is not fully tractable in scenarios where only a small number of studies are used for meta-analysis. In meta-analyses, using large numbers of studies brings into play the law of large numbers, with an expectation that estimates for the main outcome will settle around some average. With few studies in a meta-analysis, experiment factors like IFN-γ dose, treatment timing, mouse background may very well lead to meta-analysis contributions that are not solely down to study size and may not be close to an average effect of IFN-γ. For the assessments here, the only way to interpret meta-analysis results is to maintain the original biological hypothesis at very broad levels.

Mouse strains. The meta-analyses use studies based on both C57BL/6 mice and Balb/c mice. Both C57BL/6 and Balb/c are inbred strains with different properties at phenotype level, with the former susceptible to several diseases like atherosclerosis, type II diabetes or general obesity, and the latter susceptible to various tumours and as a source of monoclonal antibodies. C57BL/6 mice are assumed to be less susceptible to microbial infection than Balb/c mice(Bohn, Heesemann et al. 1994, Heinzel, Rerko et al. 1998, Leakey, Ulett et al. 1998). It has been suggested that the macrophage response to IFN-γ activation may differ between these two mouse strains (with some caveats regarding macrophage type and particular mouse strain phenotypes). Based on those studies (Oswald, Afroun et al. 1992, Dileepan, Page et al. 1995, Mills, Kincaid et al. 2000), C57BL/6 mice may be more efficient in IFN-γ induced nitric oxide (NO) production and activation through IFN-γ may activate different and opposing signalling pathways, in C57BL/6 this would be the control of intracellular pathogens (iNOS pathway) and in Balb/c this would be wound repair (arginine pathway leading to production of ornithine). It has been acknowledged that it is the IFN-γ induced production of NO that confers resistance to microbial infections to C57BL/6 mice (Santos, Andrade et al. 2006). IFN-γ inducible NO production has also been linked to JAK-STAT and ERK signalling pathways (Blanchette, Jaramillo et al. 2003). In summary, it is very conceivable that microarray samples or studies based on either Balb/c or C57BL/6 mice will have several specific differences in gene transcription in response to IFN-γ stimulation. While this would be valuable to pursue in the presence of a larger available set of studies with either mouse strain, for this thesis the meta-analysis hypothesis has to centre on the IFN-γ inducible effects that are common to these two strains, and may also be a secondary problem to that of very different IFN-γ doses and treatment durations.

Study size and number expectations. The results section uses the terms ‘larger’ and ‘smaller’ studies in context of the microarray studies used. It should be clear that while a comparison of seven IFN-γ treated samples with seven control samples is a reasonable (if not good) size for a microarray study, it is of course to be

considered a very small sample size in context of clinical trials or population studies. Although estimates for average number of studies in meta-analyses are difficult to find, it is expected that the number of studies available for meta-analysis here is also on the small side. Additionally, this meta-analysis is very unbalanced, with one study of reasonable sample size and 3 (or 5, see below) studies of questionable size for most statistics or metrics. For many areas of biological investigation, these are the facts on the ground, and make a test of pragmatic use important despite the statistical shortcomings. The RP model may benefit from the inclusion of the two smallest studies in the meta-analysis, which are too small to even obtain the necessary mean and variance estimates for parametric methods. However, out of a total of K_{tot}=71 ranked differential expression combinations, the two extra studies only contribute 3 of those. A jack-knife procedure would be a possibility for investigating the effect of individual studies, i.e. leaving one study at a time out of meta-analysis. Although the usefulness would be limited due to the small number of studies, it would conceivably clarify broad study size effects.

Multiple testing. This occurs in any statistical analysis that applies statistical inference testing to more than one variable at a time or to multiple pairwise sample group comparisons. It has the effect of increasing the chance of obtaining false positive results amongst all these tests, because the probability of making a false positive call (significance level or *alpha*) is set at a given acceptable value for a single variable (gene, in this case) or comparison and not applicable to more than one test. Simultaneous statistical hypothesis testing on multiple variables has been well understood and described for general statistical inference (Hsu 1996) and the issue has been introduced in the early days of microarray technology (Dudoit, Yang et al. 2002, Storey and Tibshirani 2003), although problems remain regarding the algorithm used for adjustment and the assumption of statistical independence between genes. Without applying any type of correction, large gene lists will contain many discoveries that are purely down to chance or rather error. If avoiding false positive results is a priority and gene-hypothesis p-values are intended for publication, multiple-testing corrected p-values need to be used, but

even then the assumed independence between variables (again, genes in this case) remains problematic, in that a particular gene target may be represented by more than one probe sequence on the array, or two genes may entirely depend on one another or are similarity affected by a third variable in their transcriptional activity. In any of these cases, the expression levels of these genes cannot be considered independent of one another. However, for machine learning approaches, network graphs or other bioinformatics approaches it can be beneficial to initially allow a higher proportion of false positive results that are subsequently removed through domain knowledge or filtering steps. This is because two genes may be similar in transcription activity even if they fall on different sides of a selected statistical criterion. The statistical methods here only use a reasonably stringent threshold ($p \leq 0.01$) for determining what constitutes a discovery and will therefore contain more false positive results than necessary. This is helpful for the biological validation in chapter 5 and it avoids having to take into account performance of very different multiple-testing correction algorithms such as Bonferroni or Benjamini & Hochberg. It should be noted that comparisons between RP and the other two models are complicated by the fact that the former uses a permutation Null distribution to obtain statistical significance, whereas the other two models use reference distributions (chi-squared for FP, standard normal for ES). In theory, the RP methods reported estimated percentage of false positive predictions is equivalent to using a False Discovery Rate p-value adjustment on p-values derived from reference distributions, or alternatively the distribution properties of ES and FP statistics could be ignored and a Null distribution generated instead.

Alternative statistical inference. Combining per-study estimates of statistical significance across all studies is also subject to the method used to obtain the per-study estimates. This is here particularly evident for the FP model using parametric (Welch t test) and non-parametric (Rank Product test) input, where the former is amongst the worst performers and the latter amongst the best. There are many other possible choices that are not addressed here. There are various updates and alternatives to the effect size estimator d in the ES model, any many other

possible parametric or non-parametric tests as input for the FP model. This does not apply to the RP model, it is based on fold-change between groups and as such does not get assigned a per-study significance-estimate, although the overall geometric mean of ranked ratios could conceivably be replaced by other point estimates. In addition to the statistical test, it would also be possible to make further suggestions for the meta-analysis models themselves, e.g. apart from different weighting schemes for the RP model, the inherent effect size model weighting could also be modified, or a more explicit (p-values themselves are already affected by study size) weighting scheme introduced for the FP model. Since the number of possible avenues of investigation would quickly multiply the necessary assessment output, only a subset of options and variations can be considered here.

Model assumptions. A large and expected part of the difference in results between the meta-analysis models is related to their underlying assumptions in order to correctly test hypotheses. Independently of which statistical test is used, FP assumes study homogeneity and is sensitive to individual biased studies in that a single large p-value can outweigh several small p-values. There are modifications like restricting per-study statistical p-values to a given threshold (e.g. 0.05), a procedure that is similar in principle to trimmed means (Zaykin, Zhivotovsky et al. 2002). However, this in itself makes the assumption that large p-values are a product of bias (e.g. publication bias, one-sided testing), which is difficult to assess and may not be true. The ES model does not assume study homogeneity, but can account for heterogeneity if there is evidence for it. It otherwise shares weaknesses with other parametric methods, in that the effect size statistic assumes normality and homoscedasticity of experiment groups in all studies. It also relies on estimates for group means and variance being available for each study, which is a potential problem when relying on published results rather than having access to the full study data. The RP model is not subject to many assumptions about underlying distributions, which makes it robust, but the absence of assumptions about the distribution of the tested parameter in a population also means that results cannot necessarily be extrapolated to population-level. For microarray

studies, log-normality of gene expression is commonly assumed, but it has already been shown (Breitling, Armengaud et al. 2004, Smyth 2005) that in context of small studies, some non-parametric methods or the shrinkage of gene expression standard errors towards a combined value (from multiple genes) improve results over parametric options. If this holds true in meta-analyses of microarrays, then RP and non-parametric FP (and models not applied here) are better options independent of distribution assumptions being correct or not.

Simulated data. It could be argued that simulation would provide more comparative information between meta-analysis performance, but it is not attempted here because trade-off between time spent and high-value results is likely to be affected by the required number of parameters and parameter values. While some of the reference articles cited do include simulation runs, the parameter range is by necessity limited and ultimately not referencing biology. For the purposes of this thesis, a full simulation would need to take into account combinations of number of studies, size of studies, quality of studies, test type, missing values, and biological effects of different treatment regimes on variance and differential expression. Given that a resource for a direct biological assessment of results was available, this is the preferred choice here.

Missing values. By removing gene measurements that are incomplete across all samples and studies, the issue is avoided entirely in this meta-analysis. This affects hundreds to thousands of gene probes and is therefore potentially detrimental to results. The issue is complex this type of missing data has not been fully addressed before and is therefore the subject of chapter 4.

Addressing the main questions posed at the outset of this chapter, the assessment and notional changes of meta-analyses yield sufficient results to answer them in numerical and broadly biological context, with the interpretation of true value in biological terms reserved for chapter 5. The answers to those four questions are provided first before extending the discussion to more detailed issues.

Does meta-analysis provide additional results when compared to individual studies? By any form of assessment, meta-analyses add value beyond individual study analysis even with small numbers of studies, small sample sizes and a range of experiment factor values. The largest study in itself provides a large number of findings present in the meta-analysis and will largely drive any meta-analysis, but additional meta-analysis results are returned.

Are there performance differences between the three meta-analysis methods? There are fairly large differences, the theoretically powerful effect size model is the worst performer, showing very little robustness to small study numbers and small study sizes. The rank product method or Fisher's combination of probabilities have very clear performance advantages, but it is important to chose the most suitable variant.

Can suitable meta-analysis alteration be identified? For the rank product model, the suggested proportional weighting scheme appears to perform at least as good and in some circumstances better than the standard model and should be considered for more thorough investigation and publication. When choosing Fisher's model, meta-analysis results clearly benefit in quality (if not in quantity) if input data sets are analysed under non-parametric assumptions or more specifically, using rank product analysis for individual studies. This cannot necessarily be generalised to other non-parametric tests.

Do meta-analysis results show biological relevance? All results show notable overlap with DPM's manually curated gold standard list of genes associated with IFN-γ in murine macrophages. Meta-analysis is on par or marginally better than analysis of a single 'good' study where the most highly significant genes (of individual studies) are concerned, but meta-analysis performs better than any single study where genes of more moderate levels of significance (in individual studies) are concerned. The highest biological relevance is achieved by RP models and the non-parametric FP model.

Chapter 4

Imputation of missing values in microarray meta-analyses

This chapter tests the possibility of replacing (imputing) special cases of missing values for gene expression data in meta-analysis data sets. Six imputation methods are applied to sets of artificially introduced missing values.

The applied meta-analyses in chapter 3 were performed on 9812 genes, ignoring more than 3000 genes that are not present on all microarray platforms. This raises the question if there is any benefit of imputing replacement values for these genes rather than accepting their loss in meta-analysis.

Imputation of missing expression data values in microarray studies is not a new subject. Many algorithms exist that aim to replace a missing expression level measurement for a gene in a given microarray sample with an estimated expression level, where this estimation is based on data obtained from non-missing data of other genes or the same gene in other samples. However, this type of imputation has not been attempted for a special case of missing values in microarray meta-analysis, where merging data from different microarray platforms can result in a multitude of missing data that are not rooted in data acquisition problems or physical chip problems, but in the chip design of different microarray platforms.

The investigation and proposed solutions are based on semi-synthetic data, making use of the same source data as used for the meta-analyses run in chapter 3 and described in detail in chapter 2. Imputation algorithm performance is tested on the basis of artificial and repeated introduction of missing values amongst the full set of known gene expression values, allowing a final comparison of imputed (estimated) gene expression with known (observed, original) gene expression.

4.1 Introduction to microarray missing value imputation

In meta-analyses, the necessity to combine data from multiple studies at the level of statistical results or other per-study metrics requires a data mapping step in which probe transcripts from one type of microarray are mapped to transcripts on another type of microarray, assuming not all studies use the same microarray platform. This can easily lead to hundreds of probe transcripts that exist on some microarray platforms, but not on others, and it can be assumed this set of transcripts is made up of both newly identified genes that were not incorporated into older chip designs, sequences that were included due to some sequence selection algorithm but may or may not have a real biological function, and some microarrays having a particular biological focus. Although this seems to suggest that unmapped probes cannot automatically be regarded as irrelevant, for the sake of simplicity or a focus on statistical methodology rather than biological results these transcripts are often explicitly excluded from meta-analysis data sets as not having complete observations (Rhodes, Barrette et al. 2002, Hong, Breitling et al. 2006, Thomassen, Tan et al. 2009) or the problem is avoided altogether by only considering studies that have been performed with the same microarray platform (Sims, Smethurst et al. 2008, Gyorffy and Schafer 2009).

General mechanisms for 'missingness' of data were first formally discussed by Donald Rubin in 1976 (Rubin 1976) and subsequent publications, defining missing-completely-at-random MCAR, missing-at-random MAR, and not-missing-at-random NMAR, all of which relate to the probability of a value missing being dependent on observed or unobserved data. In context of microarrays, most researchers implicitly or explicitly assume MAR or MCAR, where the probability for the value missing is not dependent on the value itself. This assumption is not necessarily correct in cases where background noise is higher than signal, or where values are removed by the user because they fall into an arbitrary low expression range, i.e. the probability of the observation being missing is increased because its expression value is low. In terms of investigating

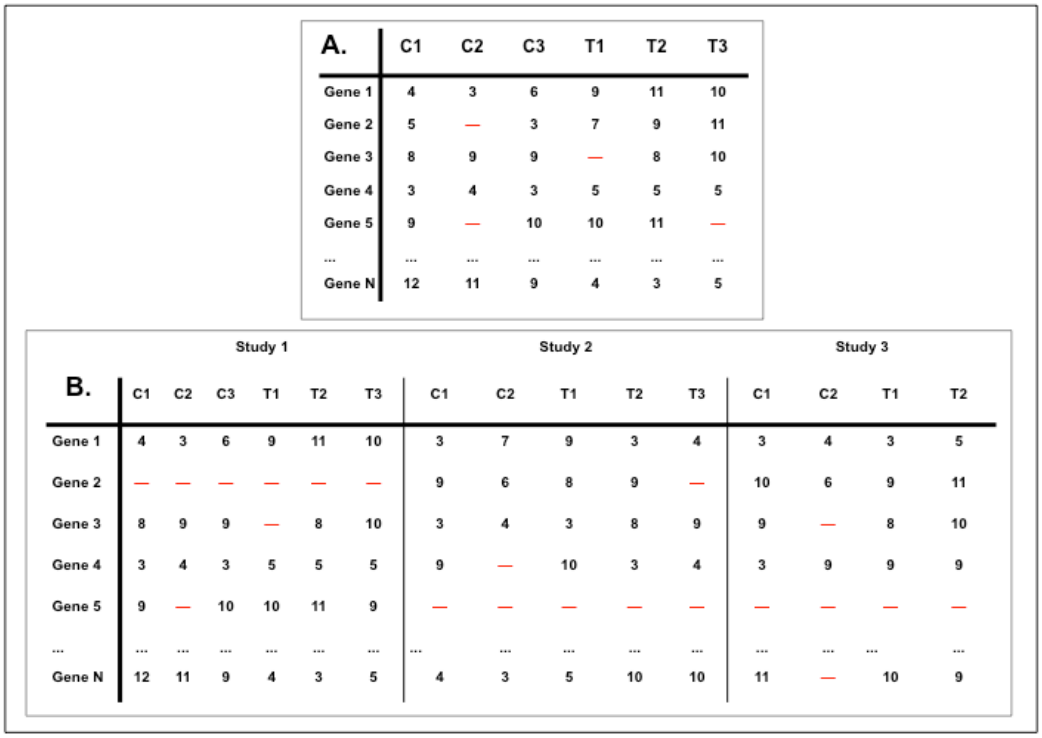
gene observations missing in a meta-analysis through platform design, the reason for given gene transcripts not being present on particular microarray platforms may differ from gene to gene or be driven by external factors such as array manufacturing cost or updates to the knowledge on known gene sequences. Imputation here is therefore on the basis of these genes being missing at random or completely at random.

In terms of missing values in microarray studies (not meta-analyses), the most commonly investigated types of missing values include scratches or dust on the array surface interfering with the laser scanning process, faulty probe deposition mechanics or spot morphology. This usually results in a small percentage (often assumed to be around 5%) of observations for a gene in a study missing, and presents a problem for those researchers using multivariate analysis approaches like Principal Components Analysis and Support Vector Machines, or generally algorithms requiring complete data matrices to work robustly or to work at all. Consequently, it was this research community that first introduced methodologies to recover or impute missing values within a microarray study, beginning with Troyanskaya proposing a K-Nearest-Neighbour gene-similarity approach (Troyanskaya, Cantor et al. 2001). This was sporadically followed by newly introduced algorithms like a Bayesian Principal Components imputation model (Oba, Sato et al. 2003) or least squares based imputation (Bo, Dysvik et al. 2004, Kim, Golub et al. 2005), and more recent attempts to focus the imputation on genes with function similarities as determined by GO gene ontology database (Xiang, Dai et al. 2008) or taking a higher-level approach in which the imputation method applied is dependent on data characteristics (Brock, Shaffer et al. 2008). It should be noted that most of these algorithms include simple reference approaches in form of replacing missing values with a constant or a sample mean, and these are invariably shown to be inferior.

These methods apply to microarray studies in general, but similar research has not been conducted for the imputation of missing values in microarray meta-analyses. Although the above concepts can of course be extended for meta-analyses by applying an imputation algorithm to each study in turn, this does not encompass a

situation with all observations missing for a gene in all samples of a microarray data set. The difference between these concepts is outlined in figure 4.1.

Figure 4.1 Missing gene expression values in microarray studies



A.

	C1	C2	C3	T1	T2	T3
Gene 1	4	3	6	9	11	10
Gene 2	5	—	3	7	9	11
Gene 3	8	9	9	—	8	10
Gene 4	3	4	3	5	5	5
Gene 5	9	—	10	10	11	—
...	...	...	...	...	...	...
Gene N	12	11	9	4	3	5

B.

	Study 1						Study 2					Study 3			
	C1	C2	C3	T1	T2	T3	C1	C2	T1	T2	T3	C1	C2	T1	T2
Gene 1	4	3	6	9	11	10	3	7	9	3	4	3	4	3	5
Gene 2	—	—	—	—	—	—	9	6	8	9	—	10	6	9	11
Gene 3	8	9	9	—	8	10	3	4	3	8	9	9	—	8	10
Gene 4	3	4	3	5	5	5	9	—	10	3	4	3	9	9	9
Gene 5	9	—	10	10	11	9	—	—	—	—	—	—	—	—	—
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
Gene N	12	11	9	4	3	5	4	3	5	10	10	11	—	10	9

Panel (A) represents a gene expression data matrix (genes in rows, biological samples in columns; C and T are control and treatment samples, respectively) obtained from a single microarray data set, panel (B) represents three aligned/mapped gene expression data matrices as used for meta-analysis of microarray studies, where alignment is based on the identity of a gene. Horizontal solid red lines indicate missing data. The difference between panels (A) and (B) lies in the nature of their missing values, with (A) missing some gene expression values due to localised array quality issue and individual studies in (B) potentially also suffering from this issue, but with the addition of genes for which there are no observations at all in a single study because its microarray chip type does not include a gene probe that is found on the microarray chip types used in the other two studies.

By extension of the above single-study approaches to imputing missing data, it should be possible to identify genes relevant to imputation of a missing gene in a study by borrowing information from other independent studies. In effect, this means treating a set of independent studies as a unified data matrix or single large study. This is conceptual rather than factual, as studies are difficult to combine at data level (otherwise meta-analyses would hardly be necessary) and they remain independent. Applying single-study imputation methodology in a multi-study

context is currently unexplored, as is the loss of information that is brought on by excluding genes from a meta-analysis because they are missing in a subset of studies, or the reduction in the power of a meta-analysis to detect overall effects across fewer studies that a gene is measured in.

The question posed here is therefore if a missing gene's expression values can be imputed for an entire study wholesale, given that there is available but independent data for this gene in other studies, and that use could be made of gene-to-gene similarities to identify suitable other variables (genes) to use for imputation. The ability of performing such an imputation also needs to be investigated for benefits compared to leaving these values missing in a meta-analysis.

The studies used in this thesis comprise two Affymetrix chips (two each), MGU_74Av2 and MOE_430a. The two platforms contain 22690 and 12488 gene transcripts, respectively. An Affymetrix-designed mapping algorithm[29] identifies an intersection set of 9812 as "good matches". If one uses the smaller microarray platform probes as reference, this means out of the potentially available 348000 data points (~12000 genes x 29 samples), 58000 could be considered missing values (~2000 x 29 samples), which is a missing value frequency of around 17%.

Where a broader range of microarray platforms is used and the number of studies is larger, it may still be possible to obtain reasonable overall effect sizes for each gene by simply accepting missing values. As exemplified by this thesis itself and the reference works cited, this availability of larger numbers of studies is not a given and meta-analysis results are therefore limited in their genome coverage. The motivation underlying the investigations in this chapter therefore lies in testing if there is an advantage in using statistical models and rules to estimate large volumes of missing data rather than accepting the loss of hundreds to thousands of observable variables (genes) from the meta-analysis results.

R scripts generated for the analyses in this chapter are contained in the electronic supplementary material ("HF" for high frequency of missing values):
ImputationMissingValues.R, ImputationMissingValues_HF.R

[29] http://www.affymetrix.com/support/technical/comparison_spreadsheets.affx#1_1)

4.2 Imputation algorithms

In order to address the question of replacing missing data with estimated data in a realistic context (with numbers of studies and study sizes small), the same data sets as used for meta-analysis in this thesis are used again, with a selection of existing or alternative imputation methods, and evaluation on a small number of outcome criteria. Assessing imputation requires comparison of imputed values to known values, which means these data cannot be used in full but are used as the basis for a limited simulation where status missing/non-missing is set rather than observed, that is, missing values are artificially introduced.

For all imputation methods, the initial design matrix is identical to that in their original use:

$$G \in \mathbb{R}^{m \times n} = G_{MN} = \begin{bmatrix} g_{1,1} & g_{1,2} & \cdots & g_{1,N} \\ g_{2,1} & g_{2,2} & \cdots & g_{2,N} \\ \vdots & \vdots & \cdots & \vdots \\ g_{M,1} & g_{M,2} & \cdots & g_{M,N} \end{bmatrix}$$

That is, gene expression matrix *G* with *m* genes and *n* samples for the original application of these methods represent a single microarray study, but for the application here the concept is expanded to including the samples of all studies. An additional study index is therefore not shown here and would only be used in some of the assessment methods (where each study's values are set to missing in turn to assess the effect study size has on imputation). Similarly, an index for treatment status is not included here, as only one method below uses treatment group specific imputation (control samples predict control samples and treatment samples predict treatment samples).

In the absence of a de facto standard for imputing missing values in a multi-study setting, several different methods were chosen for imputation of missing values.

4.2.1 KNN

KNN (Troyanskaya, Cantor et al. 2001) imputation identifies genes with expression profiles (gene *m* across samples *n*) most similar (based on K nearest neighbours algorithm) to the gene with missing values, and the imputation value is the average (weighted by similarity) expression of those similar genes in the same sample. Importantly, this approach is 'perpendicular' to BPCA and linear regression as described below, in that KNN identifies similar genes to impute missing values, whereas the other methods identify similar samples to impute missing values.

Algorithm description:

- Identify gene with missing values, the target gene
- Using non-missing data for this gene (i.e. other studies), calculate Euclidean distance between this gene and other genes
- Identify K (user-specified) nearest genes to the target gene
- Missing value for target gene is imputed as the average (weighted by distance) of the K nearest genes from the same sample
- Repeat for all missing values in samples

4.2.2 BPCA

BPCA (Oba, Sato et al. 2003). The Bayesian Principal Components algorithm has been shown to be one of the best performing imputation algorithms (Jornsten, Wang et al. 2005). It uses a principal component regression on samples with non-missing values to predict a sample with missing values, that is, missing gene expression values are estimated based on their expression in (a linear combination of) similar samples rather than estimating missing values based on the expression of similar genes. Instead of using individual samples to predict a response variable (the sample with missing values), it first reduces the set of predictor variables to its major dimensions or principal axes (in decreasing order, linear combinations of samples that best represent variance between samples). Three advantages of this

methodology are a) the use of lower-dimensional space (samples rather than genes) to perform the estimation of missing values, b) the lack of requirement for any user provided model parameters and c) it emphasizes important principal components by shrinking less important ones towards zero. The model initial state begins with missing values imputed by gene expression means across samples, a variational Bayes then both estimates the models posterior parameter distribution and re-imputes the missing values, iteratively arriving at optimal model parameters using an Expectation-Maximization algorithm.

4.2.3 Array-wise KNN

Array-wise KNN is a simple use-adaptation of KNN, based on the question of how this approach works when using similarity of samples (across all genes) rather than similarity of genes (across all samples). In order to do so, one simply rotates the data matrix through 90 degrees before applying KNN imputation to see if the relationship between sample expression vectors instead of gene expression vectors can successfully be used to replace the missing values. Imputation is therefore done via averaged expression of the same gene in the most similar arrays in other studies.

4.2.4 Array-wise linear regression

Array-wise linear regression is introduced as a simple alternative to BPCA. Like BPCA, it uses information from the same gene but other samples to impute missing values in a given sample. Instead of iterative parameter estimations and Bayesian probabilistic PCA it simply uses regression coefficients in the estimation of missing values. The linear model considers all genes in a given array as a response variable that can be predicted by the genes in all samples of all other studies. Using the coefficients obtained from this model, missing values for a given gene in a given array can then be imputed from non-missing instances of that gene in other studies, providing a weighted mean replacement.

Algorithm description:

- Identify an array (the target array) containing missing values.
- Apply a linear model where the target array is the response variable and all arrays from all other studies are predictor variables.
- Impute missing gene expression value in target array using the expression values of the same gene from all arrays in other studies, each multiplied with its prediction coefficient (intercept added subsequently). In other words, this is a weighted mean with the weight depending on the "predictive" strength of an array regarding the target "response" array.

4.2.5 Array-wise linear regression by sample type

This is identical to the linear regression above, but instead of using all arrays in other studies, it uses only predictor variables (arrays) that match the type of the response variable (array with missing values), where type is either treated or control samples. This is to investigate the balance between a more biologically targeted prediction and a reduction of statistical power because the number of available predictors is reduced.

4.2.6 Ranked sets

This scheme first identifies genes similar to a missing gene (based on other studies), then uses a weighted average of those genes to impute the missing gene's value. While the distance metric and final imputation matches that of the KNN algorithm, the main difference lies in the selection of a gene set to use for imputation. Where in KNN these are the k nearest neighbours in terms of similarity of gene expression profiles to the target gene *across* all other studies, the ranked set selection here considers the similarity of genes to the target gene *separately* for each study. Subsequently, the per-study lists are combined based on the rank product statistic. This methodology is introduced because in theory it is

independent of the linear or non-linear relationship between studies (between-study normalisation does therefore not matter).

Algorithm description:

- Identification of a gene with missing values in a study (target gene)
- For this gene in *each* of the other three studies, obtain Euclidean distance to all other genes
- In each study, rank all genes by similarity to target gene
- For each gene, compute a summary statistic across three studies, indicating how consistent its ranked similarity is to the target gene. The statistic is the rank product (same method as described in chapter 3) of each gene across three studies.
- Identify 50 genes with the highest consistency across studies in terms of their similarity to target gene
- Those genes are the selected set
- For each gene in the selected set of genes, compute a weighting factor for each study (normalised inverse Euclidean distance)
- Average weighting factors for each selected gene across three studies
- Impute value for missing target gene in a sample as the weighted mean of the selected most similar 50 genes in that sample.

4.3 Evaluation methodology

Gene selection and introduction of missing values. In order to assess the effect of a missing value imputation, the 'truth' needs to be known. Since this is not known for the actually missing gene data in these studies, semi-synthetic data are created from the original, this is here achieved by starting with a subset of genes (n=9812) that are in common to all four studies. To avoid random selection of only noise genes (which would not enable subsequent assessment of treatment effects), this is reduced to a gene set of 6564 genes that are biased towards containing a proportion of differentially expressed genes, through a requirement that each gene have at least 1 out of 29 samples with an expression value greater than $\log_2(10)$. A single random selection of 600 genes is then drawn from this set, with the rationale that this (given that 100 of these 600 are artificially set to missing)

mirrors the missing value frequency (17%) for the data sets used in this thesis. This initiation of missing values on the same set of 600 genes is repeated (randomly, with replacement) 50 times, that is, 50 different sets of 100 genes are independently set to completely 'missing'. This generates a total of 5000 imputations per method, with this allowing to establish accuracy and stability across repeated sets with respect to the identity and possible different expression characteristics of any single set of selected genes. Genes are set to missing for each study in turn, and it is always the same set of 100 genes for each study.

Imputation of missing values. For each resampling instance and gene, missing values are imputed by each of the algorithms methods listed above. Or in the comparative instance, left as missing.

Evaluation of imputation quality. Two outcomes are assessed, the first is the accuracy with which an imputation method recalls original values, the second is an imputation method's ability to recover biological and statistical differences between treatment and control groups. Accuracy is here estimated through the normalised root mean squared error (NRMSE, in this case normalisation is to the range of original values) comparing the original gene expression values (prior to their being set to "missing") with imputed gene expression values. Biological and statistical relevance is represented by $\log_2$ scale differential expression between IFN-γ and control group averages, and by Fisher's sum of logs meta-analysis (individual studies are analysed by Welsh's t test), again comparing original outcomes with outcomes on imputed data.

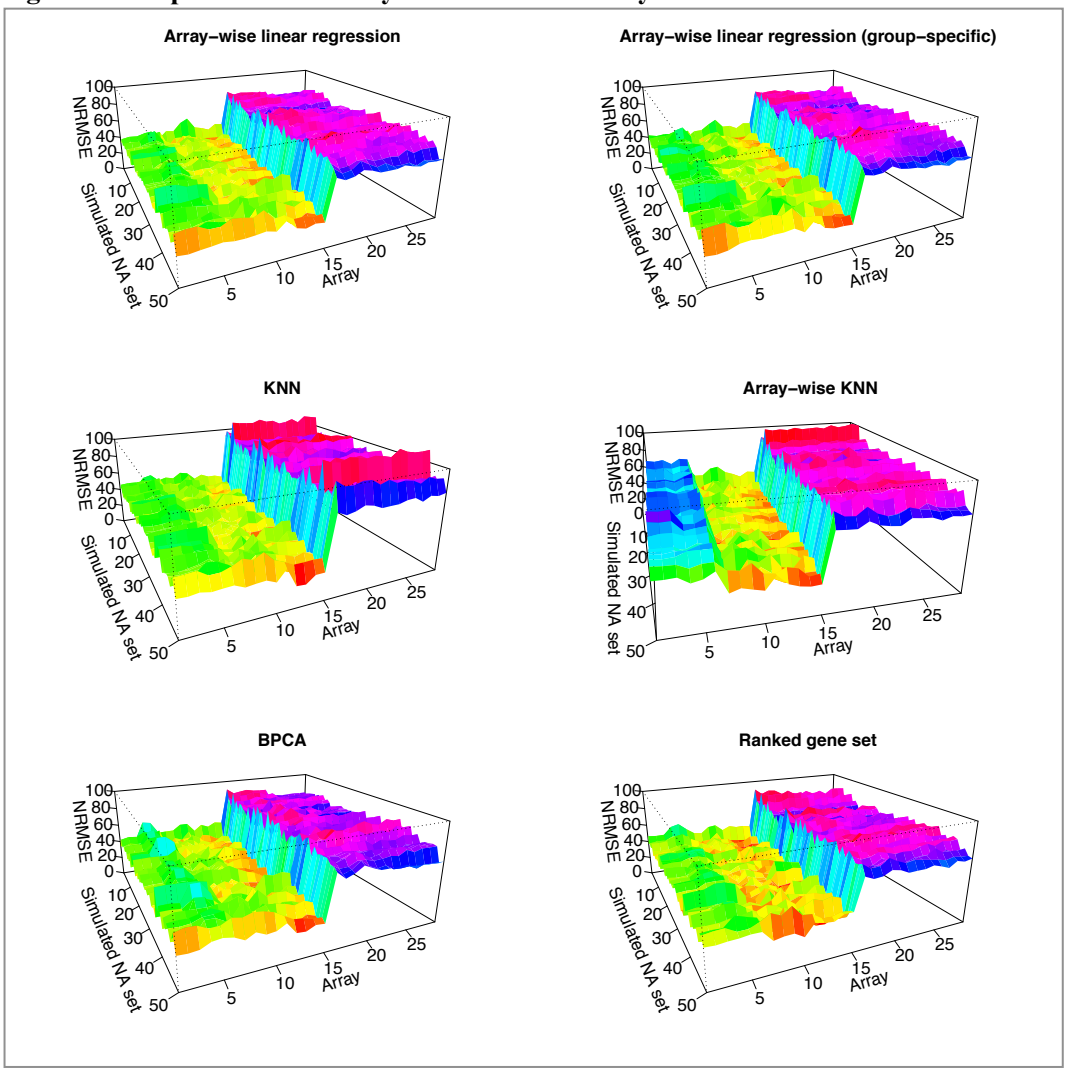
4.4 Results

Imputation accuracy depends on amount of non-missing data

There are 29 biological samples (arrays) in the four microarray studies used for meta-analysis (with 6, 5, 4, 14 samples for studies 1,2,3,6, respectively). Looking at the imputation accuracy (z axis), it is clear from figure 4.2 that the largest normalised root-mean-squared error (NRMSE) is linked to the largest study (the

right-most 14 arrays in this graph). This is most likely to be evidence for the fact that if there are missing values for the largest study, then the imputation of those missing values suffers from a lack of other available data. That is, for each array in this study only 15 other arrays (from studies 1,2,3) are available for imputing a gene's expression level. For the smallest study (study 3), each arrays' missing values can be imputed based on 25 arrays in the other three studies, increasing imputation success. This figure also highlights that in terms of error, all six imputation algorithms are broadly similar to one another, and that this error is reasonably stable across the application of those algorithms to 50 separate random sets of genes.

Figure 4.2 Imputation accuracy for individual arrays



Each graph shows the imputation error (measured by NRMSE) for comparing imputed values to (known) original values. **X-axis** ("Array") represents each of the 29 arrays

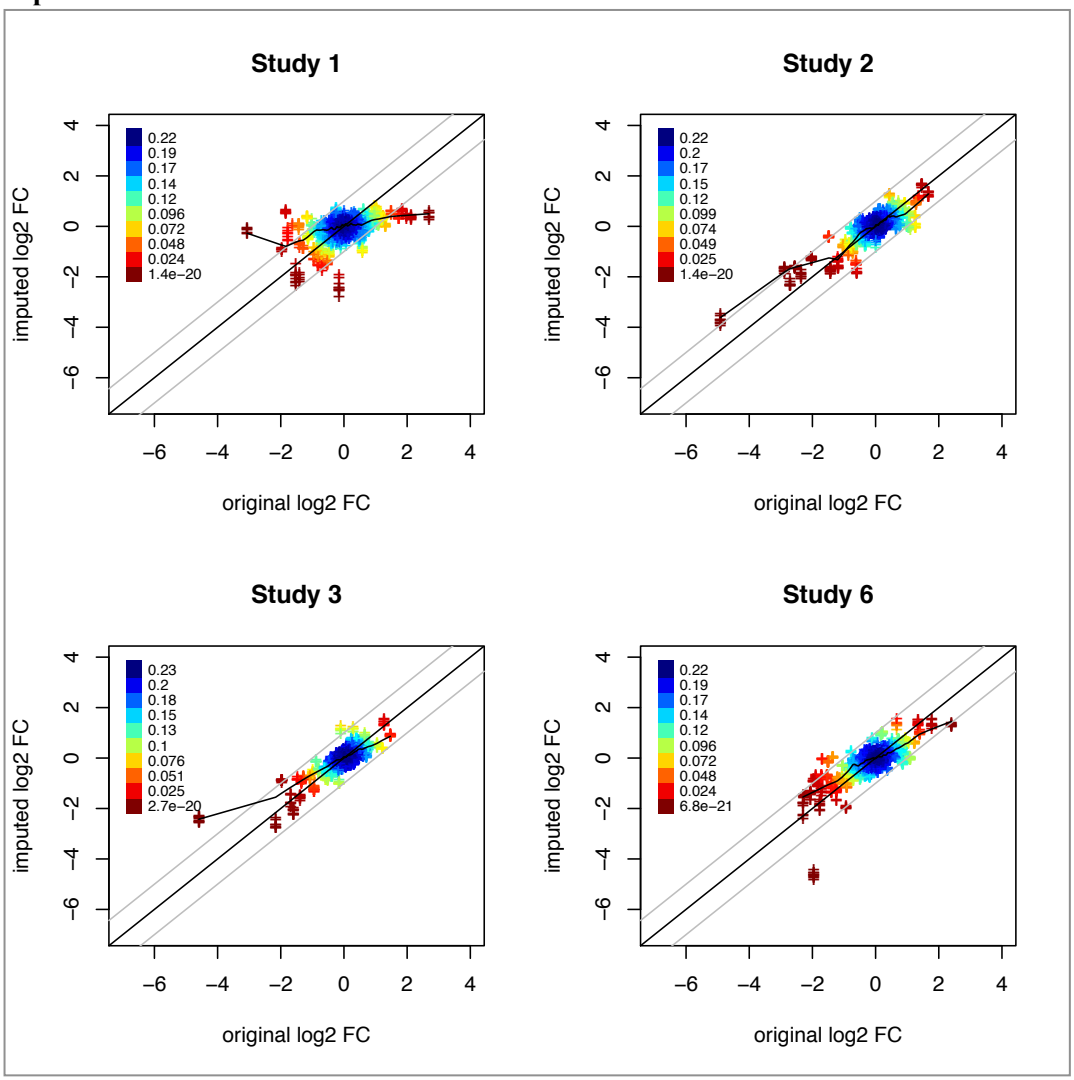
belonging to four microarray studies. Arrays 1 to 6 belong to study 1, arrays 7 to 11 belong to study 2, arrays 12 to 15 belong to study 3 and arrays 16 to 29 belong to study 6. **Y-axis** represents each of the 50 randomly sampled data sets (of 100 genes each) where the expression value was set to "missing". All random sets were drawn from a set of 6564 genes that is enriched for high expression measurements. **Z-axis** represents the normalized root-mean-squared error (NRMSE) of comparing the original expression values for 100 genes in that random set against their imputed expression values. NRMSE=0 indicates that imputed values are identical to original values.

Differential expression is not well recovered by imputation

Here, imputed gene expression values are compared to the original expression values in terms of a biological measurement. Differential expression between control and treated samples is usually measured in terms of average fold change, in case of $\log_2$ scale data, this is for each gene the difference in the arithmetic mean of treated samples compared to the arithmetic mean of control samples. A zero represents no change, a positive value up regulation and a negative value down regulation. In figures 4.3 and 4.4, the best (linear regression on arrays) and worst (KNN) differential expression outcomes are visualised by plotting the original (known) fold change against the fold change based on imputed values. Other algorithms fall in between these two and are not shown, with BPCA performing similarly to regression and array-wise KNN and ranked set selection performing similarly to KNN. A perfectly recovered fold change measurement would fall on the centre diagonal. Two features are evident from these graphs. One, the amount of data available for estimating a replacement value plays a crucial role, as can be seen by the strong deviation of data points from the centre diagonal in studies 1 and 4, suggesting that for these larger studies there are not sufficient numbers of samples from other studies to obtain a good imputation estimate. Two, simple array-wise linear regression is associated with best recovery of the original fold change values, but even this method provides at best reasonable replacement estimates for genes in the two smaller studies (2 and 3). Overall, the accuracy of imputation is insufficient for biological assessments. A two fold (+ or – 1 on $\log_2$ scale) change is often seen as biologically relevant change in mRNA levels, and imputation difference alone comes close to this for many genes and may cancel out or exaggerate biological differences, particularly where imputation is performed on the larger studies 1 and 4 (with less data available from the remaining studies for the estimation process). This imputation insufficiency is particularly notable in the KNN and the gene set selection algorithm, where imputed values are only marginally different from zero (which is equivalent to noise as the majority of genes in a study will not be differentially expressed). In a real application, with original values not known, imputation

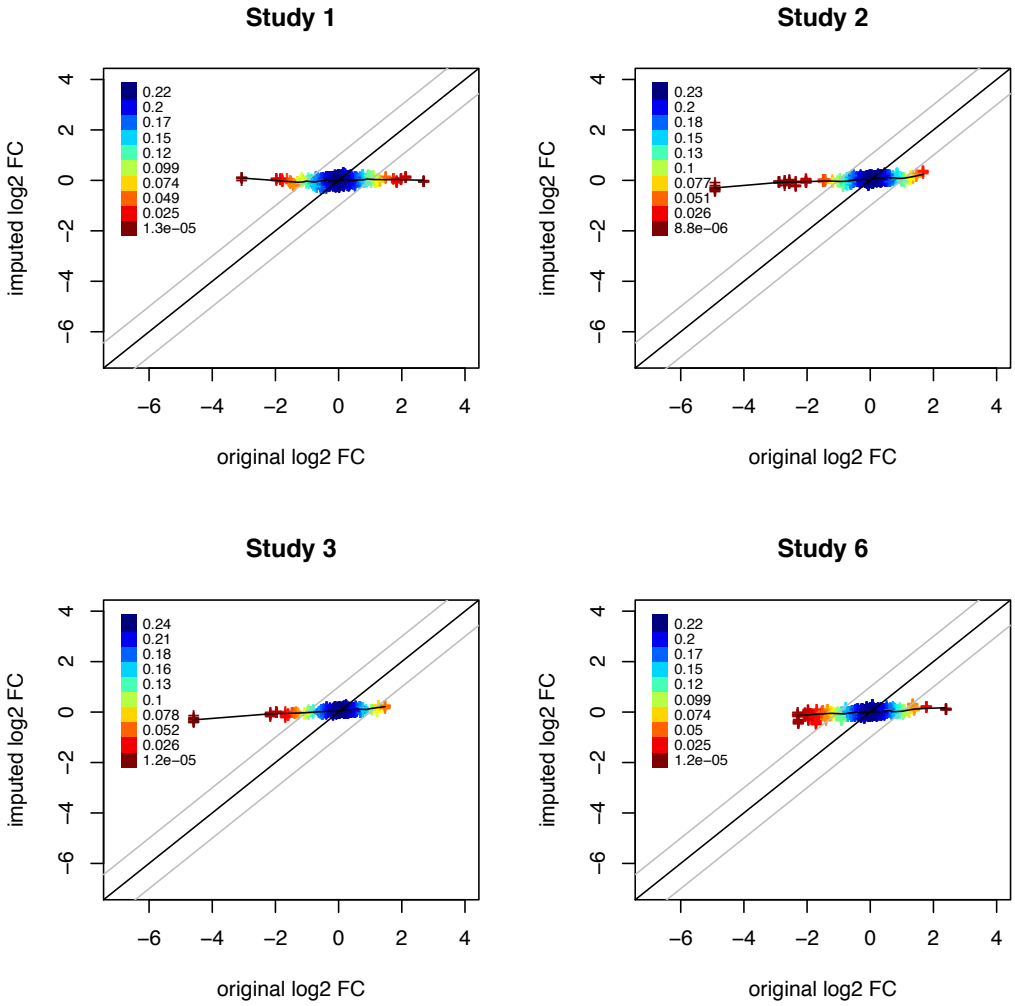
would therefore not clarify for any given gene if the observed fold change is real or due to lack of imputation quality.

Figure 4.3 Recovery of differential expression values by array-wise linear regression imputation



Each graph shows the effect (on biological fold change estimates) of imputing gene expression values that were artificially set to status "missing" in one of the four studies. **X-axis** represents the originally computed fold change between control and treated samples for 50 random sets of 100 genes (5000 data points). **Y axis** represents the same computation after artificially setting each set of 100 genes to status "missing" and imputing a replacement estimate for them based on data from other studies. Fold change values based on imputed values and matching original fold change values fall on the centre **diagonal** line, the upper and lower diagonal indicate where an imputation would be 2 times higher or lower than the original fold change. A LOWESS **line is fitted** to all data points to indicate data trends. The colour scale maps standard bivariate normal density at each point and shows the estimated density value.

Figure 4.4 Recovery of differential expression values by KNN imputation



Each graph shows the effect (on biological fold change estimates) of imputing gene expression values that were artificially set to status "missing" in one of the four studies. **X-axis** represents the originally computed fold change between control and treated samples for 50 random sets of 100 genes (5000 data points). **Y axis** represents the same computation after artificially setting each set of 100 genes to status "missing" and imputing a replacement estimate for them based on data from other studies. Fold change values based on imputed values and matching original fold change values fall on the centre **diagonal** line, the upper and lower diagonal indicate where an imputation would be 2 times higher or lower than the original fold change. A LOWESS **line is fitted** to all data points to indicate data trends. The colour scale maps standard bivariate normal density at each point and shows the estimated density value.

Imputation underestimates biological fold changes

Irrespective of good or bad performance (exemplified by figures 4.3 and 4.4), imputation consistently underestimates fold changes, that is, if the original known fold change is notably positive (up regulated genes) or negative (down regulated genes), the imputed value tends to be closer to 0 rather than a larger positive or

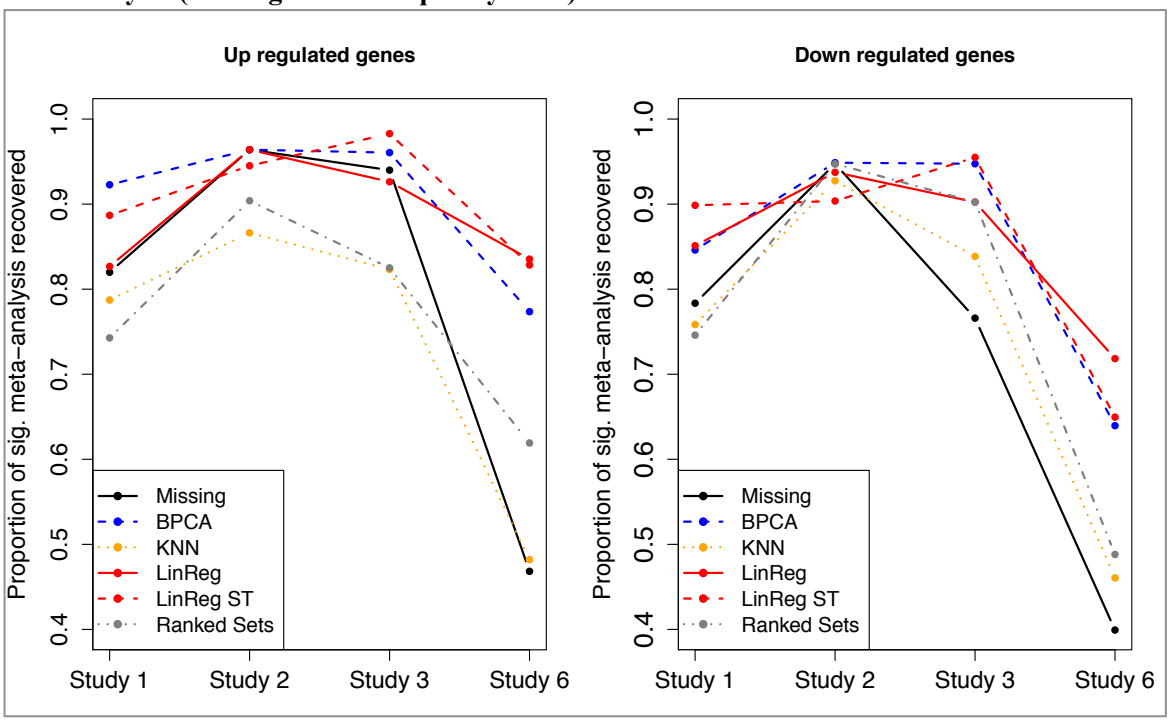
negative value. There are individual genes for which the opposite is true or where this effect is minimised in the best imputation and with sufficient amounts of data (figure 4.3, studies 2 and 3), but these are not in a majority. The effect is caused by imputation assigning similar estimates to missing values in control and treatment samples, meaning noise in form of a large number of low or non-expressed genes has a greater influence on the estimates than biologically active genes. This is not improved by two theoretically suitable modifications. The first modification (sample-type specific linear regression) is to explicitly estimate expression for missing values in treated samples through non-missing values in other treated samples, and conversely using control samples to predict control samples. The second modification (ranked set selection) uses a multi-step algorithm to curate a small set of genes with statistical similarity to those with missing values. However, in the simulation runs both modifications reduce (data not shown, but within the range established for best and worst imputation algorithm) rather than increase the quality of imputation in terms of differential expression measurements. For the modified linear regression model, it will certainly be the case that the imposed limitation of estimation only through the same sample type has a negative effect by reducing the number of samples available to the linear regression model, that is, the weighted-mean replacement estimate for a gene has less power. The outcome for the ranked set selection algorithm is almost identical to the KNN algorithm, which suggests that despite the very different algorithms, their application identifies and is limited by the same features in the data it is applied to.

Imputation recovers statistically significant results

Biological fold changes as described above are a measure of differences between groups that does not take into account within-group variation (other than the mean being affected in non-normally distributed gene expression values). From a statistical point of view, the question is if the imputed missing values recover the original within-group variance levels and are therefore able to identify statistically significant gene expression differences between control and treatment groups. The logical follow-up question is if this imputation performs better than simply

accepting missing values in the meta-analysis. This concerns the trade-off between reduced statistical power (with a reduced number of studies available for inclusion) and imputation error. With the original data, meta-analysis by Fisher's sum of logs results in 583 out of 5000 genes (50 random sets of 100 genes each) significantly up regulated and 799 genes down regulated at $p<=0.01$. Taking this to be the standard, figure 4.5 shows the proportions of these genes that are still identified as significant when meta-analysis is applied with data missing, or with missing data imputed from non-missing data. Three notable results are illustrated. One, BPCA imputation comes closest to the original meta-analysis results, but remains similar to linear regression. Two, meta-analysis with missing values is generally inferior to meta-analysis using imputed values, although this is less pronounced for missing values in smaller studies (with more samples available for imputation). Three, all methodologies are dependent on the size of the study in which gene expression values are missing. This is of course most notably for study 6 (which contains 14 out of 29 samples), reducing imputation effectiveness when keeping missing values.

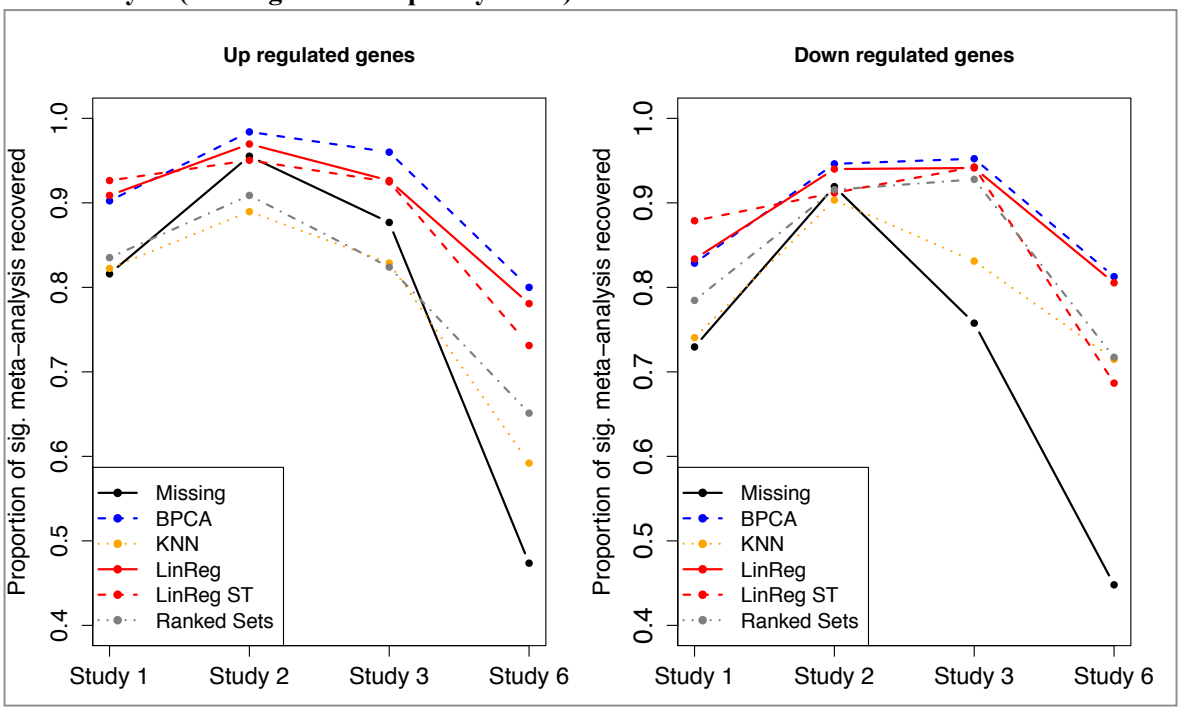
Figure 4.5 Comparison of meta-analysis on missing or imputed data to original meta-analysis (missing values frequency 17%)



This graph identifies the proportion of genes that a Fisher's sum of log meta-analysis on missing or imputed data can identify as statistically significant, where a proportion of 1 means the method detects 100% of the same genes as the meta-analysis on the original data does. **X-axis** signifies the study for which data was artificially introduced as missing. **Y-axis** signifies the proportion of statistically significant genes that are in common with the results obtained on the original data (583 and 799 for up and down regulation, respectively). "Missing" refers to meta-analysis where missing values were not imputed but left missing. "BPCA", "KNN", "LinReg", "LinReg ST" and "Ranked Sets" refer to meta-analysis performed using BPCA, KNN, linear regression, linear regression by sample type and ranked sets imputation, respectively.

The above analysis is based on the frequency of missing values matching that inherent to the original microarray studies for this meta-analysis (17%). While not covering a full spectrum of missing value frequencies, another statistical assessment was run at a missing value frequency of 1.5% (100 genes out of 6564 per array) in order to determine if the imputation quality is strongly dependent on this. Figure 4.6 suggests that there are only small differences in the number of successfully retained meta-analysis results between having 500 or 6464 non-missing genes available for estimating replacement values.

Figure 4.6 Comparison of meta-analysis on missing or imputed data to original meta-analysis (missing value frequency 1.5%)



This graph identifies the proportion of genes that a Fisher's sum of log meta-analysis on missing or imputed data can identify as statistically significant, where a proportion of 1 means the method detects 100% of the same genes as the meta-analysis on the original data does. **X-axis** signifies the study for which data was artificially introduced as missing. **Y-axis** signifies the proportion of statistically significant genes that are in common with the results obtained on the original data (583 and 799 for up and down regulation, respectively). "Missing" refers to meta-analysis where missing values were not imputed but left missing. "BPCA", "KNN", "LinReg", "LinReg ST" and "Ranked Sets" refer to meta-analysis performed using BPCA, KNN, linear regression, linear regression by sample type and ranked sets imputation, respectively.

4.5 Discussion

Unlike missing value imputation in individual microarray studies, the imputation of values missing due to different probe coverage of multiple combined microarray data sets has not been investigated before. With some justification, this can be assumed to be because the generation of mRNA expression values for large numbers of genes that have not even been included on a microarray chip does seem to implicate fabrication of data. However, this has to be weighed against the major concern of losing thousands of genes in a meta-analysis because they are not represented on all microarray chip platforms. With large numbers involved, a

simple investigation as described in this chapter is therefore used to give this problem numerical consideration instead of accepting intrinsic assumptions.

Developments in missing value imputation. Since the time the work in this chapter was carried out, other algorithms have been devised to impute missing values in microarray studies, and at least one convenient online tool (MissVIA[30]) has been developed to run many of these on a user-supplied data set. Of the algorithms described in this chapter, it includes KNN and BPCA. While this is highly convenient, it is aimed at imputation within single microarray studies characterised by a different and infrequent type of missing values.

Normalisation. Referring to normalisation across all studies (only within-study normalisation has been carried out), it is assumed that data are not normalised (where this means mathematical matching of per-array gene expression distributions between all arrays) across studies. In case of BPCA, this is explicitly stated as preferable to normalisation, which the authors' empirical evidence has shown to degrade imputation performance. For KNN, it is not stated if normalised or non-normalised data are preferable, but it is here assumed that different studies are only systematically and linearly different from one another. This means that if one study provides higher expression level readings than another, these higher levels are consistent for most genes and have less effect on measuring similarity of genes or samples than it would have on measuring statistical differences (where a 5% higher expression level for all genes in an array would affect a statistical metric). This assumption would extend to arrays within a study, but is here not necessary because individual studies are normalised as they were for meta-analysis in chapter 3.

Differential expression vs. statistical significance. Notably, if assessment of imputation is limited to differences between group mean expression levels, results would be limited and, in terms of differential expression, underestimated (figure 4.3). This applies to even the best methodologies (BPCA and array-wise linear

[30] http://cosbi.ee.ncku.edu.tw/~godkin/

regression) and is indicative of biological fold change being an ill-defined criterion with no really meaningful threshold. With imputation causing a decrease (rarely an increase) of observed average fold changes, genes would easily be discarded as not biologically relevant, and it is only due to the better defined thresholds of statistical testing that an advantage of imputation over keeping missing values is established. After imputation, a gene may be associated with a less significant p-value, but for a large proportion of genes these p-values still are significant at the same allowance level for false positive results (figure 4.5). For fold changes, no such objective cut-off exists and an analyst would likely consider many different and arbitrary fold change thresholds in a trial-and-error approach to identify genes of interest.

Frequency of missing values. Statistical results for data missing with high frequency (17%) are almost identical to results from data missing with low frequency (1.5%), as demonstrated in figures 4.5 and 4.6. These two global missing values frequencies do not cover more complicated patterns of missingness (such as data missing for more than one study, or additional data missing at random or not at random) and are therefore not entirely conclusive without further study. However, the current computation are similar to what imputation in single microarray studies achieves (Scheel, Aldrin et al. 2005), with imputation in their assessment losing around 6-15% of statistically significant results.

Influence of study size. It is clear from figure 4.5 that imputation of missing values depends on study size. The more samples are available for estimation of a replacement value, the better the imputation. This means that if values for the largest study (study 6) are missing and being imputed, statistical power is limited by the lower number of available samples in the other three studies. However, even in this circumstance, imputation (particularly using BPCA and linear regression, which recover around 80% of the original results) is considerably more effective than performing meta-analysis with data for this study remaining missing (which recovers only 47% of original results). The study size effect is therefore clear, but not a counter indication to imputation.

Simulation and parameters. The promising and not entirely expected results in this chapter may warrant a more expansive assessment. This chapter assesses the value of imputation only within the real data as used for this thesis as a whole. If this pragmatic data selection had proved ineffective for imputation of missing values, alternative simulation models would not provide much confidence in their applicability. Given that imputation shows some advantages over keeping or ignoring missing values, a more wide-ranging simulation could determine more exact parameter boundaries. This will likely prove necessary for purposes of publication, because the imputation of complete observations for a gene in a microarray study is entirely different from established methods for infrequently and randomly missing data. A full simulation would in particular address the effect of a range of study sizes, study numbers, gene expression levels, missing value frequencies and missing values patterns. With a large parameter space to cover, the results of the current work would be useful – apart from motivating further investigation - in limiting the number of imputation approaches to BPCA and linear regression.

Gene selection and resampling procedure. There are alternative implementations to assure stability of imputation estimates with respect to the number and expression properties of gene chosen as missing or non-missing. The current implementation involves repeated resampling of the artificial introduction of missing values to one fixed set of 600 genes, with each sample setting a different set of 100 genes to completely missing. It would also be possible to repeatedly sample gene sets of size 600 from the list of 6564 genes (biased for high expression), with a subsequent single introduction of 100 missing genes to each set. It would also be possible to do both in combination or to vary the number of non-missing genes while keeping the proportion of missing genes stable at 17%. The here favoured option is an attempt to sensibly limit the parameters of investigation to the crucial stage, that is, the artificial introduction of missing genes where the non-missing genes may consequently have different expression characteristics depending on the selection.

Imputation through BPCA and linear regression of samples. It is perhaps not surprising that linear regression on samples performs similarly well to BPCA imputation method, as both are based on similarity of samples. For microarrays, missing value imputation methods often make use of the correlation structure available in these high-dimensional data with few observations (samples) and large numbers of variables (genes). However, in context of missing values for all observations of a gene in a study, it appears to be not the gene correlation structure (as used by KNN and ranked sets) but the sample correlation structure that informs the best results. KNN is often assumed to perform less well than BPCA (Jornsten, Wang et al. 2005) unless there is a strong 'local' correlation structure in the data (strong gene-gene similarities). Based on the observation that linear regression is conceptually similar to BPCA and also performs similarly, the idea was tested that imputation could be further improved by only using samples of the same class (treated or control) for the linear regression model, that is, if a missing gene's observation is in a treated sample, only treated samples are used to estimate the replacement value. This does not improve imputation quality here, presumably because any additional value gained is lost through lowering the sample numbers for estimation.

Ranked set gene selection. The newly introduced ranked set selection algorithm shows few advantages over the KNN algorithm, itself shown not to be a good method in the context of this chapter. Given the similarity of outcomes, the newly introduced method evidently only takes a different algorithmic path to arrive at very similar imputation estimates to KNN. This may be due to relationships between studies being sufficiently linear to provide no noticeable advantage to the per-study approach used in the ranked set method when compared to the across-all-studies approach used by KNN. In terms of gene selection schemes, both of these two algorithms are inferior to BPCA and linear regression, which suggests that a) these studies do not have a strong gene correlation structure that would help gene-correlation based algorithms, and b) imputation improvements for the ranked set selection would have to be directed towards including biological information,

that is, genes not only similar in expression profiles in a study or meta-analysis, but also known to have functional association through other sources of information.

Strength of evidence versus scope of proposed procedure. What is attempted in this chapter is fundamentally different from the usual imputation of missing gene expression values in microarray study. The proposal here is that all observations for a gene in a complete microarray study can be generated in their entirety, that is, a gene is "invented" from scratch where there has originally not been a gene. However, three factors make this scenario worthwhile exploring in microarray data. One, it is known which genes form part of the genome and therefore should be represented on a microarray platform; where an "invented" patient is fictional, an "invented" gene is at least known to exist. Two, the high-dimensionality of microarrays means there are a very large number of variables available with which to predict missing values. Three, a meta-analysis that omits (and excludes from analysis) genes missing on a particular array type is likely to miss out on a large number of reasonable biological results. While these factors motivate the investigation in this chapter, the obtained evidence in favour of imputation of complete genes cannot yet be considered sufficient to match the scope of the chapter. Next steps will likely require a full simulation of imputation parameters and a biological assessment of genes identified through meta-analysis of incomplete or imputed data.

Conclusion. Surprisingly, the imputation of data for genes entirely missing from a microarray platform shows clear advantages over the alternatives of keeping missing values, or omitting genes from analysis altogether. Despite differential regulation being somewhat underestimated after imputation, the number of significant meta-analysis results matches that of the original data upwards of 80%. Any reduction in the actual p-value is less important than maintaining a constant type I error level, and better than no results at all. It is also clear that imputation strategies based on sample-correlation (BPCA, linear regression) outperform those based on gene-correlation (KNN, Ranked Sets). While this is promising, it cannot

be taken as complete verification of imputing the type of missing data discussed in this chapter. Firstly, all results are based on semi-synthetic data derived from a single meta-analysis set of studies. Secondly, only a narrow parameter space is tested, leaving unclear how higher frequencies of missing values, availability of more studies or samples, or different gene-correlation or sample-correlation structures in microarray studies could affect these outcomes.

To date, missing value imputation appears not to have been tested for anything other than individual microarray studies with randomly missing data values, and given the positive outcomes in this chapter, this supports publication in this space with the above necessary additions to the work in this chapter.

Chapter 5

Biological context of meta-analysis results

5.1 Introduction

This chapter develops results obtained through meta-analysis of IFN-γ microarray studies into biological context, with the aim of exposing new biological insights.

Interferons are classically separated into three types. Type I interferon signalling is characterised by interferon alpha or beta (or all with the exception of IFN-γ) proteins binding to the interferon alpha receptor on the cell membrane and activating the JAK-STAT pathway, resulting in STAT1/STAT2 heterodimers that can form a complex with IRF9 (then a heterodimer), translocate to the nucleus and activate genes containing ISRE (Interferon Stimulated Response Element) motifs in their promoter. Type II interferon signalling is typified by IFN-γ binding to an IFN-γ receptor and also activating the JAK-STAT pathway, but where this results in the formation of STAT1 homodimers that translocate to the nucleus and activate genes containing GAS (Gamma Activated Sites) motifs in their promoter. Type III interferon signalling is less well characterised, and is assumed to share its function with type I interferon, but with JAK-STAT activation through IL28 or IL29 (and other cytokines) binding to an interferon lambda receptor. Both type I and type II interferon play a role in the activation of macrophages, although type II is regarded as the classical factor in this.

Current biological knowledge related to the IFN-γ signalling pathway can be separated into three semantic classes. One is the canonical view of the type II interferon activated JAK-STAT pathway as described above. The second class is then the set of interferon-stimulated genes (ISGs) transcribed by STAT1 and further downstream. This much larger set of genes is less well defined than the canonical pathway, but still contains what is considered as known in the cell

response to interferons in general. Notably, research suggests that this list has considerable overlap with genes activated through the type I interferon pathway. The third class contains genes for which there exists research or hypotheses that mean they may be part of the response of interferon but are pending further confirmation. These three classes are outlined and used for reference later in this chapter.

Contextualising the meta-analysis results requires both definition of gene sets related to the above three classes, as well as a condensed definition of meta-analysis test results. The full process described in this chapter can be outlined as follows. Results of the meta-analysis of multiple microarray studies investigating the effect of IFN-γ on murine macrophages are reduced to "meta-analysis-only discoveries", those genes that are statistically significant ($p<=0.01$) in a meta-analysis but not statistically significant in the analysis of any individual microarray study. It is important to note that this does not constitute the full measured type II response, because genes statistically identifiable as significant in any single microarray study are excluded. Next, result lists are reduced to a consensus list of all genes that are statistically significant in at least two out of three meta-analysis model families.

Subsequently, these results are evaluated for their known inclusion in interferon responses through an independent database (Interferome.org) of known interferon-stimulated genes (ISGs), experimental data provided by the Division of Pathway Medicine, simple gene annotation and information obtained from GeneCards[31], and searches of literature through PubMed[32] (with search term "interferon" combined with official gene symbol).

The biological context revealed in the overlap between meta-analysis results and existing research of the type II activated immune response is then assessed and novel findings presented.

[31] www.genecards.org
[32] http://www.ncbi.nlm.nih.gov/pubmed

5.2 IFN-γ signal transduction

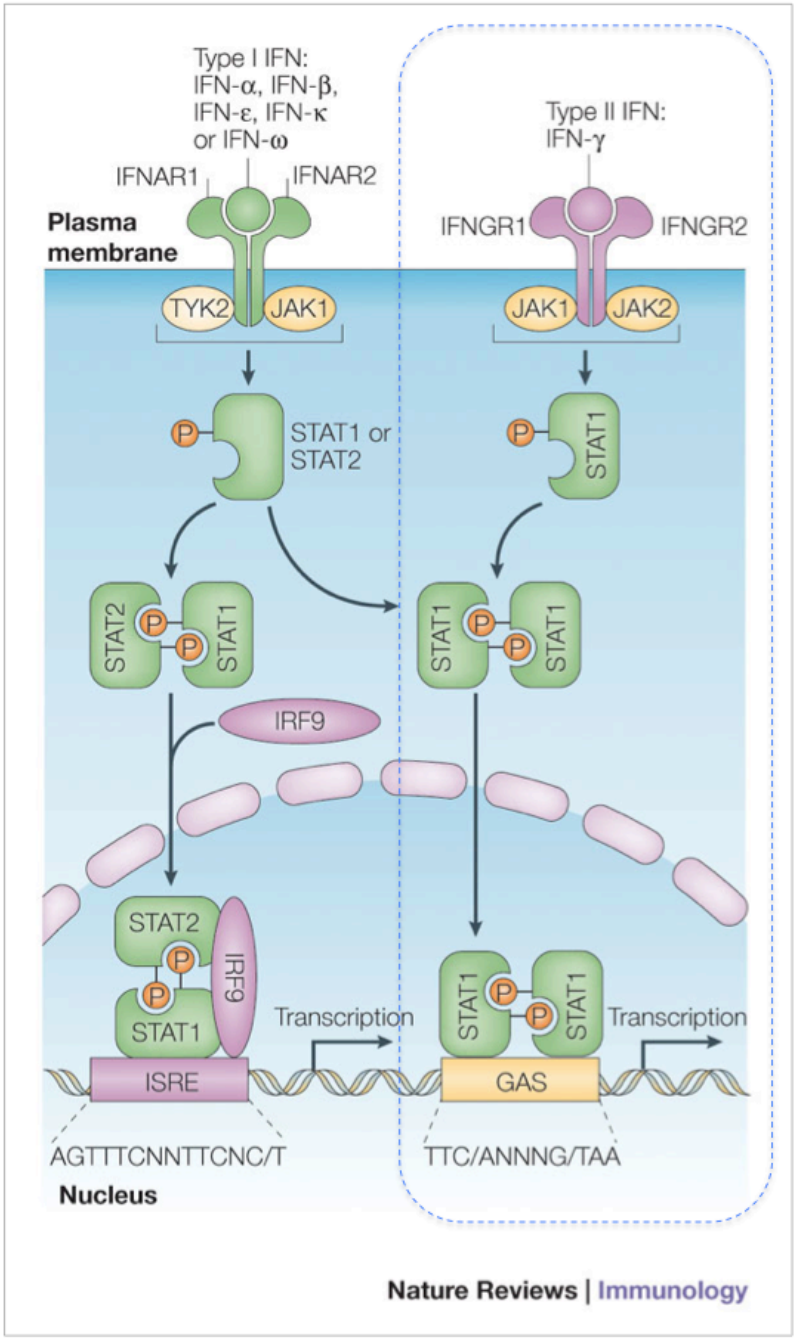
The JAK-STAT pathway plays a critical role in initiating both the innate and adaptive immune responses upon infection.

The classical activation of macrophages requires stimulation by an extracellular IFN-γ signal, which engages a cells' JAK-STAT pathway with the specific outcome of phosphorylating STAT1 proteins, which in turn dimerise, translocate to the nucleus and activate the transcription of genes containing a GAS (STAT1 homodimer binding site) motif in their promoter region. With no particular focus on the type of interferon involved, these genes and those affected indirectly are referred to as Interferon Stimulated Genes (ISGs). This paradigm is frequently represented as a pathway schematic with differences in the level of detail and biological specialisation. Figure 5.1 represents the simplest form of this pathway, with alternative schematics available from many sources (Vilcek 2003, Schroder, Hertzog et al. 2004, Platanias 2005, Horvath 2013). This canonical model is however a simplification that ignores other contributors and modifiers to the type II immune response:

- Other genes or miRNAs positively or negatively regulating (via respective proteins) the JAK-STAT pathway, such as SOCS, PIAS (Greenhalgh and Hilton 2001) or miR-155 (Lu, Thai et al. 2009) targeting and potentially down regulating SOCS (at least in T-cells), or nitric oxide synthase 2 (NOS2) associating with IFNGR1 in bacterial infection (Velez, Hulme et al. 2009). Other examples are *Prn*, *Ptpn1*, *Cd45*.
- Complicated crosstalk with other signalling pathways like MAPK signalling or PI3K-AKT (Rane and Reddy 2000), particularly type I immune response and in newer research, with *Ch25h* in sterol metabolism (Blanc 2013).
- Activation/suppression of gene transcription that is not due to STAT1 alone (such as activation of *Irf1* only through joint activity of STAT1 and TNF**α**-activated *NFκb* (Ramana, Gil et al. 2002). Other examples are *Jun*, *Cebpb*, *Hmga1*, *Myc*.

- Presence of a gamma-activated site in a gene may need to be complemented by both co-activators with STAT1, e.g. *Brca1* (Ouchi, Lee et al. 2000) and other transcription factors like SP1 bound to other sites on that gene (Ramana, Gil et al. 2002).
- In addition to interferon (or IFN-γ) related genes that are directly transcribed by STAT1, there are of course also genes (indeed, the majority of the interferon response) activated or suppressed further downstream and therefore only indirectly regulated by STAT1.

Figure 5.1 Canonical IFN-γ activated JAK-STAT signalling



[Source: (Platanias 2005). Mechanisms of type-I- and type-II-interferon-mediated signalling. Leonidas C. Platanias.; licence for reuse obtained: 3140331400980)
(additional dotted frame added for highlighting type II activated JAK-STAT pathway)]

Extracellular IFN-γ binds to an IFN-γ receptor complex (IFNGR1 and IFNGR2) on the cell surface. This activates (through autophosphorylation) the receptor-associated kinases JAK1 and JAK2, which in turn regulate STAT1 phosphorylation, allowing homodimerisation of STAT1 proteins and translocation to the cell nucleus, where STAT1 is able to bind to the promoter regions of genes containing a gamma activated site (GAS) and enables transcription of those genes until STAT1 activity is inhibited through dephosphorylation.

5.3 Known interferon stimulated genes (ISGs)

This thesis assesses the effect of IFN-γ treatment on gene transcription, it is therefore necessary to relate meta-analysis findings to those genes that are already known to be stimulated by interferons (specifically, type II). This serves two purposes, the first is simple confirmation that a reasonable proportion of results are biologically relevant rather than chance, and the second is to identify novel participants in the response.

Many research groups have compiled existing experiment-derived knowledge in the form of ISG (interferon stimulated genes) lists or databases (Boehm, Klamp et al. 1997, Der, Zhou et al. 1998, de Veer, Holko et al. 2001), where de Veer et al have also provided an early example of database based on compiled microarray results [33] . A more recent and supposedly more complete resource is Interferome.org [34] (Samarajiwa, Forster et al. 2009). Interferome is manually curated from publicly available microarray data sets (36 to date) to consolidate type I, type II or type III interferon regulated genes. The interferome study set is heterogeneous with regard to organism, cell/tissue types and type/subtype of interferon, resulting in a broad source of information that in this case may be helpful in highlighting meta-analysis-only discoveries that otherwise show little activity in individual small specialised microarray studies. This is the main database used in this chapter for validating meta-analysis results. However, the following limitations apply: 1) Due to experiment outcomes being determined by parameters like cell type, time, intervention, environmental factors, laboratory protocols, organism and of course quality of experiment design, it is not specific to any of these parameters. 2) It is unfortunately currently not possible to download the interferome data set in its entirety or to submit more than 100 ENSEBML IDs at a time (this presumably is as intended, and direct inquiries to the authors on this

[33] Database available at: http://www.lerner.ccf.org/labs/williams/xchip-html.cgi

[34] http://www.interferome.org/ Samarajiwa, S. A., S. Forster, K. Auchettl and P. J. Hertzog (2009). "INTERFEROME: the database of interferon regulated genes." Nucleic Acids Research **37**: D852-D857.

subject have not received a reply). 3) In contrast to the meta-analyses undertaken in this thesis, this resource is a simple catalogue of aggregated findings, conditional on genes being >= 1.5 fold up/down regulated, and not based on statistical models combining the underlying data.

As the Interferome database does not use any of the microarray studies that were considered for the meta-analyses described in this thesis, it provides a reasonably unbiased independent validation source against which meta-analysis results can be examined for known interferon (and IFN-γ) related genes.

5.4 Other evidence of ISGs

While the Interferome database is used as the primary resource for defining current knowledge on interferon stimulated genes, two other sources were considered for further validation purposes, one based on in-house experimental data, one based on promoter predicted GAS binding sites.

5.4.1 Experimental evidence for ISGs provided by DPM

The Ghazal lab has a research emphasis on the interaction between host and virus infections observable in macrophages, and for this purpose a large microarray study (referred to as MITCH12 and described in detail in chapter 2) was carried out to measure changes in gene expression levels across twenty-five 30-minute intervals following addition of IFN-γ to murine bone marrow derived macrophages. This study has a single replicate (treatment of BMDM culture) per time point and is therefore not amongst the set of microarray studies used for meta-analyses conducted for this thesis, providing an independent and nuanced (compared to snapshots of IFN-γ treatments versus control samples) experimental validation resource. Although it lacks statistical power through the absence of replication at each time point, its relatively frequent sampling time points and the theoretical absence of biological inter-subject variation provide useful longitudinal information in form of trends. In this case an online database of all gene expression profiles (of fully processed data) for this study is available via

GEViSE[35]. The limitations of visually classifying expression profiles into up and down regulation are described later on in this chapter.

5.4.2 Promoter predicted GAS binding site evidence for ISGs

Since the IFN-γ activated JAK-STAT signalling pathway results in STAT1 homodimers that bind to a GAS motif in the promoter regions of target genes, a succinct list of all genes with promoter predicted GAS binding sites was generated via oPOSSUM v3.0[36]. These represent all theoretical targets directly transcribed by STAT1, with this of course not covering activation or inhibition of genes further downstream. The process and parameters are described in detail in chapter 2.

Conversely, the full list of meta-analysis results was also supplied to the oPOSSUM tool in order to test the up and down regulated gene sets for enrichment with regard to other transcription factors.

5.5 Meta-analysis result gene sets

The use of six different meta-analysis models identifying separate result lists for up and down regulated genes produces twelve separate gene lists. An individual assessment of their biological context even in compressed form would cause the generation of numerous and large tables, and while this would be the most complete and detailed approach, it would detract from an assessment of how much a microarray meta-analysis confirms or adds to biological knowledge, irrespective of quality differences between meta-analysis models. This motivates an aggregation of these lists into a single gene set (although still divided into up and down regulation). Following characterisation of the original gene lists, the aggregation process is described and discussed.

[35] http://oriol.gti.ed.ac.uk/GEVisE/

[36] http://opossum.cisreg.ca/cgi-bin/oPOSSUM3/opossum_mouse_ssa

5.5.1 Meta-analysis results prior to aggregation

Prior to discussing the output of an aggregate list of meta-analysis "hits" or "discoveries", a distinction is made between meta-analysis "model" and "model family". This is based on most meta-analysis models not being fully independent of another, for example Rank Product meta-analysis: while these are three meta-analyses that all use the same analytical framework of permuted rank-based statistics and therefore constitute a "model family", the individual models work with different parameters, in this case each Rank Product meta-analysis model applies different weights to individual microarray studies. In case of meta-analysis based on effect size, only a random effect model was shown to be suitable and there is therefore no distinction between model family and model.

For any of these lists, it must be emphasized that they do not contain all possible genes statistically identified to be differentially regulated under IFN-γ treatment, because most of these genes can readily be identified in the statistical analysis of individual studies. While those genes are of course biologically interesting (given that they will be associated with larger expression fold changes), the purpose of this chapter is to identify genes that a meta-analysis can identify beyond that set.

Table 5.1 summarises the twelve original gene lists and various forms of overlap between these models and model families. It covers genes that have been identified as statistically significant ("discovered") through meta-analysis alone, i.e. this does not include genes that were discoverable through statistical analysis of any individual study. After reduction of list to genes that were annotatable and not redundant in Ingenuity Pathway Analyzer, the table lists the number of genes discovered by each meta-analysis model or model family, those unique to a model (the number of genes identified with each model that were not identified through any of the other models) or model family, and those common to all models or model families.

Table 5.1 Summary of meta-analysis-only discoveries

a) Genes up regulated by IFN-γ treatment

	Rank Product family			Fisher family		Effect size
	RP standard	RP prop. weights	RP equal weights	Fisher based on t test	Fisher based on RP test	Random Effect Model
N	70	107	75	80	72	76
Model family	133 (29)			135 (17)		76
Unique to model	10	6	11	32	1	25
Unique to model family	49			34		25
Common to all models	4					
Common to all model families	24					

b) Genes down regulated by IFN-γ treatment

	Rank Product family			Fisher family		Effect size
	RP standard	RP prop. weights	RP equal weights	Fisher based on t test	Fisher based on RP test	Random Effect Model
N	66	158	145	175	74	121
Model family	207 (38)			234 (15)		121
Unique to model	7	20	39	86	0	41
Unique to model family	112			87		41
Common to all models	3					
Common to all model families	18					

For each meta-analysis model and model family, these tables count the number of statistically significant meta-analysis gene-probes for genes up-regulated (a) or down-regulated (b) by interferon gamma. Table columns identify a given meta-analysis model and model family, table rows identify the method of counting. Counts shown exclude significant meta-analysis results that are already significant in individual microarray studies. Bracketed counts for model family give the intersection, i.e. genes in common to all models in a given model family.

Table 5.1 describes differences in gene list sizes and gene list overlap depending on meta-analysis model and model family used. Two main features are the considerable differences between the gene list sizes under different meta-analysis

models and the difference between the relatively large number of genes identified by individual meta-analysis models (or model families) and the considerably smaller number of those genes that are in common to all models or model families). Both features confirm that considerable quantitative differences exist and that they may be related to the quality of a meta-analysis. This provides confirmation that the choice of meta-analysis matters and that it is therefore useful to work with aggregated ("consensus") rather than individual lists as described in the next subsection.

5.5.2 Consensus lists of meta-analysis results

Several options for list aggregation were considered and discarded before final use of the two-out-of-three model families criterion. A union set of all genes that are identified by any of the meta-analysis models would be too large (214 up regulated genes and 392 down regulated) and include too many false positive results identified by lower quality meta-analysis models. An intersection set of genes common to all twelve meta-analysis models or common to all 3 meta-analysis model does create a valuable high-confidence set of results (see table 5.2) but is also limited to the lowest common denominator in terms of performance of meta-analysis models. Using only result lists from the best-performing meta-analysis model or model family would not be a fully objective set in that chapter 3 identified no single model clearly superseding all others, resulting in good results ignored in favour of the best results.

Table 5.2 High-confidence gene lists

Up regulated in IFN-γ (n=24)		Down regulated in IFN-γ (n=18)	
Cd47	16423; Rh-related antigen, integrin-associated signal transducer	**Degs1**	13244; degenerative spermatocyte homolog 1
Mertk	17289; c-mer proto-oncogene tyrosine kinase	**Sirpa**	19261; signal-regulatory protein alpha
Ktn1	16709; kinectin 1	**Map2k3**	26397; mitogen-activated protein kinase kinase 3
Igbp1	18518; immunoglobulin (CD79A) binding protein 1	Csnk1e	27373; casein kinase 1, epsilon
Atad1	67979; ATPase family, AAA domain containing 1	Orc5	26429; origin recognition complex, subunit 5
Slc33a1	11416; solute carrier family 33 (acetyl-CoA transporter), member 1	Fam134b	66270; family with sequence similarity 134, member B
Mrpl13	68537; mitochondrial ribosomal protein L13	Reep5	13476; receptor accessory protein 5
Fbrs	14123; fibrosin	Wdr48	67561; WD repeat domain 48
Abcc5	27416; ATP-binding cassette, sub-family C (CFTR/MRP), member 5	Cyb5r3	109754; cytochrome b5 reductase 3
Dnajc13	235567; DnaJ (Hsp40) homolog, subfamily C, member 13	Bzw1	66882; basic leucine zipper and W2 domains 1
Tspyl1	22110; testis-specific protein, Y-encoded-like 1	Ppan	235063; peter pan homolog (Drosophila)
Tmem168	101118; transmembrane protein 168	Naa10	56292; N(alpha)-acetyltransferase 10, NatA catalytic subunit

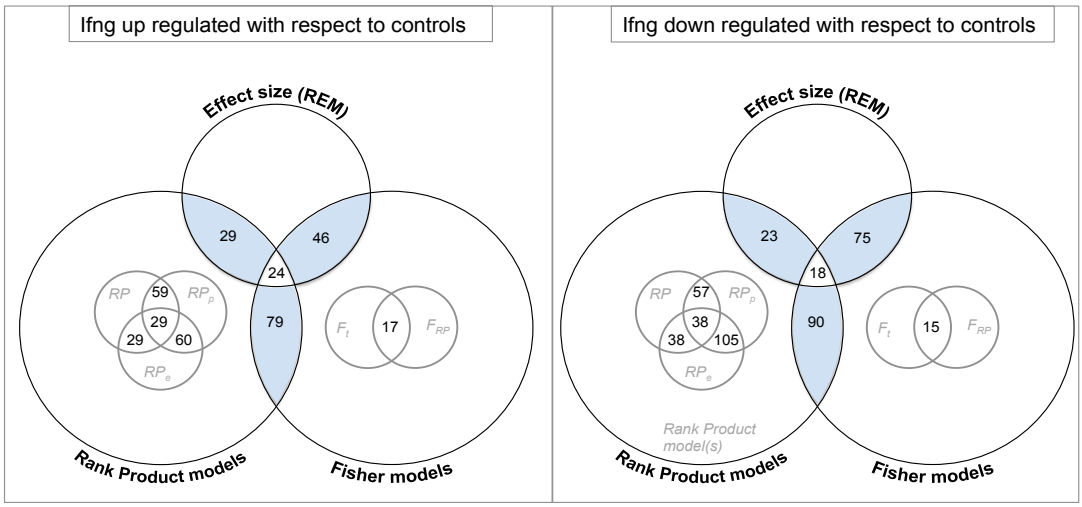
Klf3	16599; Kruppel-like factor 3 (basic)	Arpp19	59046; cAMP-regulated phosphoprotein 19
Cldnd1	224250; claudin 25	Pgrmc1	53328; progesterone receptor membrane component 1
Raf1	110157; v-raf-leukemia viral oncogene 1	Sptan1	20740; spectrin alpha, non-erythrocytic 1
Txndc17	52700; thioredoxin domain containing 17	Eif2s3x	26905; eukaryotic translation initiation factor 2, subunit 3, structural gene X-linked
Tox4	268741; TOX high mobility group box family member 4	Upf3b	68134; UPF3 regulator of nonsense transcripts homolog B (yeast)
Tulp3	22158; tubby-like protein 3	Elk3	13713; ELK3, member of ETS oncogene family
Etf1	225363; eukaryotic translation termination factor 1		
Slc35b1	110172; solute carrier family 35, member B1		
Gosr2	56494; golgi SNAP receptor complex member 2		
Tbk1	56480; TANK-binding kinase 1		
Ccdc164	381738; dynein regulatory complex subunit 1		
Ly6a	110454; lymphocyte antigen 6 complex, locus A		

Genes contained in this table are meta-analysis "hits" in all 3 model families, meaning that even the worst performing meta-analysis model is able to identify these as statistically

significant. The second column for up or down regulated genes contains the EntrezGene ID and gene description. Genes highlighted in bold can be considered as even higher confidence, as these are meta-analysis hits in all twelve individual meta-analysis models. A large disadvantage in focusing all interpretation on only this gene set is the limitation to genes that even the worst performing meta-analysis model or model family can identify, making this a lowest-common-denominator set.

While the genes presented in table 5.2 are clearly worth following up, they are also limited to the results of the least well performing meta-analysis. It therefore becomes necessary to define another alternative. This alternative is referred to as "consensus set" and is simply defined as those genes that are identified as meta-analysis-only discoveries in at least two out of the three model families. To provide an example, this would include a gene that is contained in any of the Rank Product meta-analysis family and is also contained in one of the two Fisher models or in the single effect size model. This approach assumes that qualitative differences between model families are larger than those between the models within a model family. It also allows for one model family to perform worse than the other two without affecting the aggregated result. The Venn diagrams in figure 5.2 outline details of consensus sets generation, which are all three blue-shaded intersections combined (but only counting each gene once).

Figure 5.2 Venn diagram defining consensus gene set



Large circles refer to meta-analysis model family. Smaller circles inside large circles refer to individual models within the model family (where applicable). Only intersection counts are included, totals can be found in table 5.1.
Consensus set: the union of blue shaded areas indicate selected set of genes, they are identified by meta-analysis only in any two of the three different model types. In set theory terms: $(RP \cap F) \cup (RP \cap ES) \cup (F \cap ES)$
Where Ifng is up regulated with respect to controls, n=106 genes comprise the consensus set.
Where Ifng is down regulated with respect to controls, n= 152 genes comprise the consensus set.

After this aggregation, there are n=**106** genes in the consensus set up regulated by IFN-γ, and n=**152** down regulated. These were stored in a simple data sheet[37] for the purpose of combining them with their known biological context as described in sections 5.3 and 5.4. The proteins associated with these genes can also be seen in a STRING interaction map in the appendix.

5.5.3 Meta-analysis disagreement lists

With the consensus list approach emphasizing what meta-analysis models have in common in terms of performance beyond individual study analysis, it is conceivable that some important differences between meta-analysis methods are overlooked. Whilst not discussed in this thesis, meta-analysis results unique to

[37] Available in supplementary material folder "MetaAnalysisResults", file "CategorisedCumulatedGeneList.xlsx"

each of the twelve models were also extracted (matching the totals shown in line "Unique to model" in table 5.1) for separate investigation and candidate selection at a later stage and are only included as spreadsheets in supplementary material here[38].

5.6 Meta-analysis results in biological context

Using meta-analysis consensus lists as described above, this section highlights biological context identified with reference to JAK-STAT pathway, known IFN-γ regulated genes, genes with predicted STAT1 binding sites (GAS motifs) in their promoter, relevant annotation and literature, and in-house experimental data. This supporting evidence is then combined into master lists in section 5.7.

5.6.1 In the context of JAK-STAT pathway

Only two canonical genes (including SOCS, PIAS and IRF modifiers in this definition) related to the IFN-γ activated JAK-STAT pathway are contained in the meta-analysis consensus results (*Junb, Sfpi1*). Given that these consensus lists explicitly exclude any genes that are easily identifiable (due to more pronounced changes in transcription levels on IFN-γ treatment) in the statistical analysis of individual microarray studies, this is an unexpected finding. The presence of these two genes is discussed at the end of this chapter. Most weight in terms of biological context of meta-analysis findings must therefore lie with genes stimulated by interferon (gamma), interactions with other signalling pathways or currently unconfirmed subtle modifiers of IFN-γ signalling.

[38] Available in supplementary material folder "MetaAnalysisResults", files "UniqueResultsPerModel_down.txt" and "UniqueResultsPerModel_up.txt"

5.6.2 In the context of known ISGs

Using the interferome database as reference, a simple table of known interferon stimulated genes (see table 5.3) shows that approximately 25% (up regulated) and 18% (down regulated) of genes identified in meta-analysis are already known. If specifically limiting the overlap to known type II response genes, this drops to 12% and 9%, respectively.

Table 5.3 Meta-analysis results compared to Interferome.org ISGs

Consensus list	In Interferome	Not in Interferome	Total
Up regulated in IFN-γ treatment	**26** (I=25; II=13; III=3)	**80**	106
Down regulated in IFN-γ treatment	**27** (I=27, II=14, III=2)	**125**	152

"In Interferome" identifies genes that are categorised as Type I, Type II or Type III interferon regulated genes in the Interferome database. Numbers listed in brackets show the breakdown by type of interferon, with categories not being mutually exclusive.

Making an assumption about full genome size (~20000 genes) and factoring in that the Interferome database defines 10% of these (~2000 genes) as regulated by interferome (of any type), results in table 5.3 exceed (at 25% and 18% of genes) random chance selection of IFN-γ regulated genes. This difference in observed proportion (21% if ignoring direction of regulation) and expected proportion (10%) is statistically significant at p=4.458x10^{-8} (by χ^2-test comparing two proportions), of course with the important caveat that the above assumptions are not independently confirmed. Given that the meta-analysis results intentionally exclude any easily discovered responses to IFN-γ (captured in individual microarray study and not used here), this outcome still lends some support to the relevance of the full list of results.

Meta-analysis results are also compared (table 5.4) to *in-silico* predictions of STAT1 target genes obtained through oPOSSUM. Of the 106 IFN-γ up regulated genes and 152 IFN-γ down regulated genes, 99 and 148 respectively have a unique

and valid Ensembl gene ID (required by oPOSSUM). Of these, 29 and 28 genes respectively contain a STAT1 binding site in their promoter region (when limiting searches to 5kb/2kb upstream/downstream). Under the same assumptions as for table 5.3, the proportion of meta-analysis results with GAS motif (57 out of 247 = 23%, irrespective of direction or regulation) is at p=2.831x10^{-11} significantly larger than the proportion for the whole genome (10%, although the strength of the assumption of GAS motifs amongst the 2000 interferome database genes is not known). While this appears similar to the comparison against ISGs in the interferome database, the actual gene lists differ. Of the 29 genes with a STAT1 binding site (GAS motif) in interferon down regulated genes, only 4 genes (*Gps1, *Capn2, Hey1, *Tshb*) are a known ISGs according to the interferome database. Of the 28 genes with a STAT1 binding site in IFN-γ up regulated genes, 9 are known ISGs (**Birc2, *Junb, *Rgs14, Arl6ip5, *Hnrnph2, Pnpla2, Ubr4, *Bcl3, Lpp*). Of these, even fewer (see starred entries) are known to be stimulated via a type II response. In addition, when comparing in-silico predictions of STAT1 transcribed genes to the gene expression profiles of an independent time course study (MITCH12), very few predictions are confirmed for the down regulated set (4 out of 29 have a notably decreasing expression level), but half are confirmed for the up regulated set (14 out of 28 have notably increasing expression levels). Given that the GAS motif relates to genes transcribed by STAT1 rather than inhibited, this results may to some degree validate the in-silico approach, although it is far from conclusive.

Table 5.4 Meta-analysis results compared to in-silico ISGs

Consensus list	GAS motif	No GAS motif	Total
Up regulated in IFN-γ treatment	**28**	**70**	98
Down regulated in IFN-γ treatment	**29**	**120**	149

”GAS motif” identifies genes that contain a binding site for STAT1 in their promoter region (searches include regions 5kb/2kb upstream/downstream). Up/Down regulated represent statistical results obtained through meta-analysis (and condensed to consensus lists). Row totals do not match actual meta-analysis list sizes because of identifier

mapping issues (oPOSSUM requires Ensembl IDs that were not part of the original annotation used).

Conversely, if oPOSSUM is used to identify enrichment for any transcription factors in the list of meta-analysis-only discoveries, STAT1 is featured but does not rank as one of the statistically best matches. A summary figure relating significance to GC content in sequence is included in supplementary material, a ranked list of significant transcription factors is provided in list form here (table 5.5), limited to those matching the default opossum significance threshold.

Table 5.5 Transcription factors in meta-analysis results

Fisher score rank	Significant transcription factors within set of genes identified as **up** regulated by IFN-γ	Significant transcription factors within set of genes identified as **down** regulated by IFN-γ
1	SPIB	ELK1
2	ELF5	SPI1
3	MZF1_1-4	GABPA
4	Mycn	ELF5
5	ZEB1	Zfx
6	Klf4	MZF1_1-4
7	Nkx2-5	YY1
8	Anrt::Ahr	ARID3A
9	MZF1_5-13	FEV
10	GABPA	SPIB
11	SPI1	Esrrb
12	USF1	ZEB1
13	Nkx3-2	Arnt::Ahr
14	Myc	CEBPA
15	ELK1	Stat3
16	AP1	Klf4
17	FEV	NF-kappaB
18	ZNF354C	EBF1
19	MAX	RELA

While many transcription factors are ubiquitous in their range of functions, a cursory examination of table 5.5 shows several transcription factors important to MAPK signalling (SPIB, Myc, ELK1), JAK-STAT signalling (GABPA, SPI1) or other immune related function (RELA – part of NF-kappaB complex, AP1 – cytokine response, Ahr::Arnt – inhibition of inflammatory response, ZEB1 – repressor of IL2 expression). However, a simple enrichment analysis for

transcription factor binding sites cannot equate the presence of a binding site with their biological role (where the GAS promoter is only the primary event upon STAT binding, resulting in expression of other proteins like IRF3 that in turn cause secondary events by binding to IRF promoters etc.). It is therefore unclear how useful this particular outcome is in identifying new biological context for meta-analysis findings.

5.6.3 New candidate ISGs based on existing annotation

This category covers genes that meta-analysis of multiple microarray studies identifies as regulated with statistical significance, but are not identified by interferome.org and may either have some relevant annotation or literature associated with them, or have other experimental evidence behind them. As described earlier, annotation information was obtained through GeneCards.org and literature searches. For each gene, an annotation category was subjectively decided on as follows:

Category 1. "Known Ifn related". Is the gene already known to be related to interferon regulation? This information was obtained from the interferome database, see 5.2.1.

Category 2. "Potential Ifn related". Is there otherwise any explicit mention of interferon or associated genes/processes in the pathway (e.g. Stat1, NFκb, immune response, …) of any form that could mean it could be relevant to interferon regulation?

Category 3. "No obvious relation to Ifn". Does the gene have annotation that is not obviously related to the previous two categories?

Category 4. "Lack of annotation". Does the gene have no or very little annotation?

With regard to existing annotation for genes not already known to be ISGs, table 5.6 suggests a further 12 (9 for up and 3 for down regulation) genes as ISGs, based on existing annotation.

Table 5.6 New ISGs suggested by existing annotation

Consensus list	Known Ifn related	Potential Ifn related	No obvious relation to Ifn	Lack of annotation	Total
Up regulated in IFN-γ treatment	**26**	**9**	**48**	**23**	106
Down regulated in IFN-γ treatment	**27**	**3**	**89**	**33**	152

"Known Ifn related" indicates that a gene has been flagged as type I, II, or III interferon regulated in the Interferome.org database. "Potential Ifn related" indicates that a gene's existing annotation contains explicit or implicit hints to involvement with interferons. "No obvious relation to Ifn" indicates that a gene has existing annotation but this does not appear to be related to the interferon pathway". "Lack of annotation" indicates that there is virtually no existing annotation for a gene.

The candidates in question are listed here, including some of the relevant annotation:

Bcl2a1 (direct transcription target of NFκb in response to inflammatory mediators)

Casp12 (stimulated by IFN-γ in fibroblasts, may reduce cytokine release, involved in reducing activity of NFκb)

Cd47 (inhibits cytokine production by mature dendritic cells, may be activated by IFN)

Igbp1 (depletion enhances induction of Stat1 dependent genes)

Ikbke (regulates balance between type I and type II interferon responses)

Itga4 (induced by IFN-γ, part of MAPK signalling)

Raf1 (key component of MAPK cascade, involved in macrophage proliferation and activation, affected by bacterial host infection)

Tbk1 (activates interferon regulatory factor 3, Irf3)

Txndc17 (modulates TNFa signalling and NFκb activation)

For IFN-γ related gene down regulation it is 3 genes:

Lrrfip1 (regulates TLR signalling)

Map2K3 (IL-12 induced IFN-γ activation)

Mapk1 (transcriptional repressor of interferon signalling)

An unresolved complexity in establishing known ISGs is of course that the validation set used here is based on microarray transcription studies. This means that in some cases the "new" ISG candidates are already implicated as ISGs or modifiers of IFN-γ signalling, through proteomic or other work. Nonetheless, here they are identified through microarray transcription studies and appear to be influencing or be influenced by IFN-γ signalling, but are currently not confirmed as transcription-derived ISGs.

5.6.4 New candidate ISGs based on time course data

For the purposes of validating meta-analysis candidates against experimental evidence, gene expression profiles in a microarray time course ("MITCH12") were classified by their overall expression trajectory across the full 12 hours, with each category chosen as described in table 5.7.

Table 5.7 New ISGs suggested by IFN-γ microarray time course study

Consensus list	Rising	Dropping	Peak or trough	Unremarkable	Not Found	Total
Up regulated in IFN-γ treatment	44 (42%)	**0**	**13** (12%)	**33** (31%)	16 (15%)	106
Down regulated in IFN-γ treatment	9 (6%)	**16** (11%)	**39** (26%)	**62** (41%)	26 (17%)	152

Categories are subjectively defined as follows: **Rising** (expression profile trajectory over most of the 12 hours goes from low expression to high expression, where the minimum differential between low and high is ~2 fold and does not include profiles within signal noise levels below ~100). **Dropping** (opposite criteria to Rising). **Peak** (expression profile trajectory rises and drops considerably once over the course of 12 hours). **Trough** (expression profile trajectory drops and rises considerably once over the course of 12 hours). **Unremarkable** (gene expression never changes across the 12 hours, or fluctuates without a clear pattern, or has single time-point peaks or troughs only, or is entirely within an arbitrarily defined region of signal noise below 100). **Not Found** (the gene does not appear

to be represented on this microarray platform). Numbers and percentages should be considered tendential rather than precise, because the categorisation of expression profiles is subjective.

With hit confirmation rates of 54% and 43%, respectively (combining all profiles that are not unremarkable or not found), there appears to be a reasonable amount of congruency between meta-analysis results and independent microarray time course. This is particularly reinforced by the matching regulation direction, with IFN-γ up regulated genes in the meta-analysis results primarily matching rising profiles in the time course experiment, and vice versa.

While all genes in the first three categories constitute potential candidates, amongst them are those with a weaker justification and those that are notable examples in both up and down regulation in terms of large systematic changes (generally larger than two-fold, and at higher and therefore more reliable levels of measurement) in the microarray time course that were also (and only) identified by meta-analysis of the two-condition microarray studies.

From the above set, the most (visually and subjectively) notable genes **up regulated by IFN-γ** and amongst the meta-analysis results are listed below, with highlighted gene symbols indicating the strongest expression profiles. A starred entry indicates this gene does not feature in the list of known ISGs in the interferome database and as such has higher value for follow-up as a novel candidate.

**Ahrr* (aka *Cyp1a1*, involved in cholesterol/sterol/lipid synthesis)
Arl6ip5 (regulated by type I IFN, according to Interferome.org)
Bcl3 (regulated by type I/II IFN, according to Interferome.org)
**Capza2* (interacts with *Calm1*, which is known to interact with type I IFN)
Cd47 (Receptor for SIRPA, binding to which prevents maturation of immature dendritic cells and inhibits cytokine production by mature dendritic cells. May be involved in membrane permeability changes induced following virus infection)

Cflar (regulated by type I IFN, according to Interferome.org)

Clic4 (nuclear translocation regulates macrophage deactivation)

**Ehd1*

Junb (regulated by type I/II IFN, according to Interferome.org)

Ly6a (regulated by type I IFN, according to Interferome.org)

**M6pr* (high degree of co-localisation with STAT1 and HPIV1 C proteins, but not STAT2)

Mafk (regulated by type I IFN, according to Interferome.org)

Mertk (important role in inhibition of TLR-mediated innate immune response by activating STAT1; regulated by type I IFN, according to Interferome.org)

Ms4a4b

Mxd1 (regulated by type I IFN, according to Interferome.org)

**Rnf14* (gene promoter contains Nfkb transcription factor binding site)

Rgs14 (regulated by type I/II IFN, according to Interferome.org)

Rsad2 (regulated by type I/II/III IFN, according to Interferome.org)

Sfpi1 (regulated by type I IFN, according to Interferome.org)

Stard3 (involved in cholesterol traffic and steroid synthesis)

**Usp12*

Conversely, there are much fewer (visually) very notable genes **down regulated by IFN-γ:**

Cdt1

**Csnk1e* (may act as a negative regulator of circadian rhythmicity by phosphorylating PER1 and PER2)

Elk3 (regulated by type I/III IFN, according to Interferome.org)

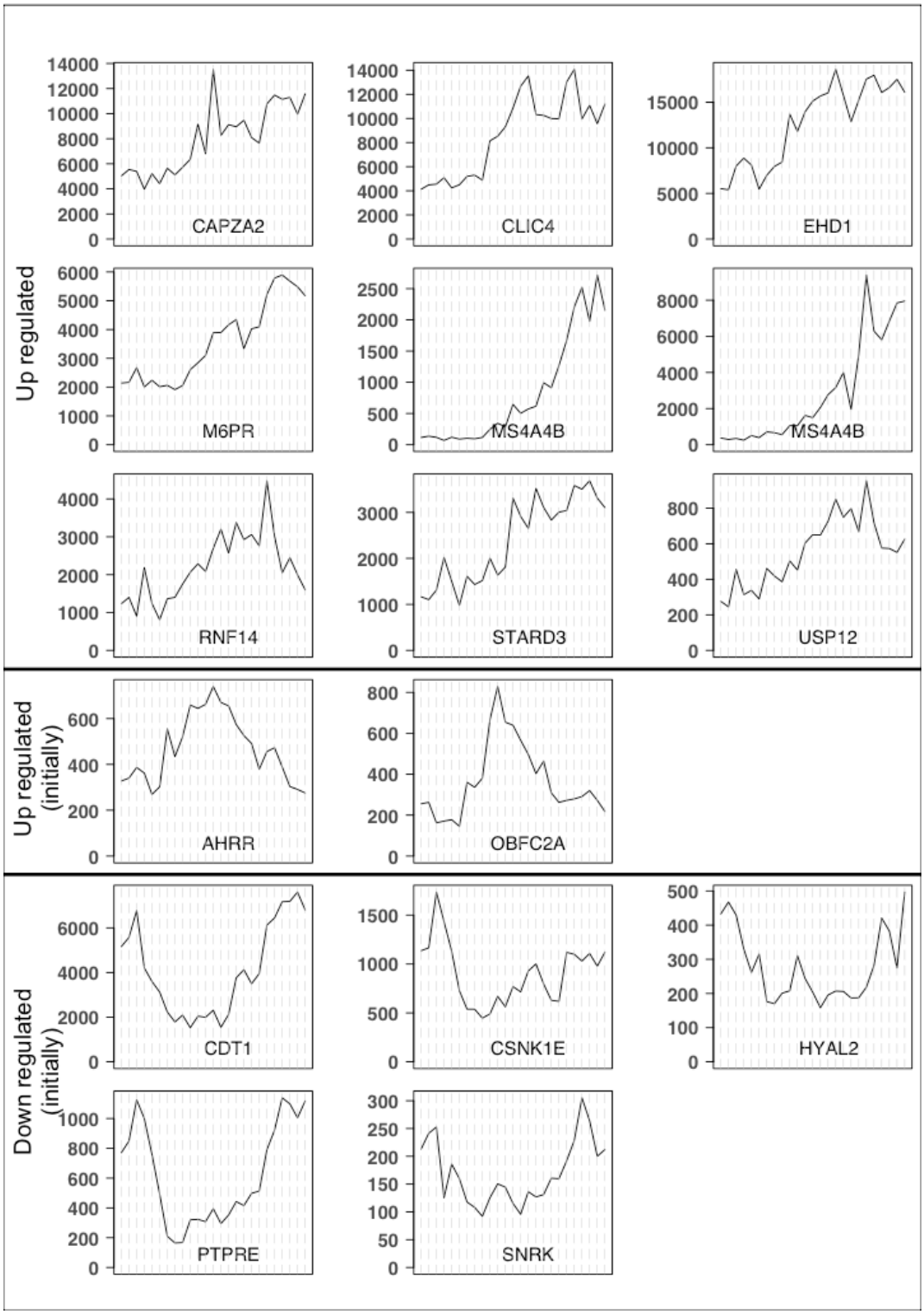
Hyal2

Ptpre (overexpression suppresses IL-6 and IL-10-induced JAK-STAT signalling)

**Snrk*

In terms of formulating hypotheses for the addition of novel candidate ISGs to already known biology, the most promising are those genes identified as significant by meta-analysis, exhibiting a notable gene expression profile over time, and not already known as an ISG. From the above list, these are ***Clic4, Ms4a4b, Stard3, Ahrr, Capza2, Ehd1, M6pr, Obfc2a, Rnf14, Usp12, Cdt1, Hyal2, Ptpre, Csnk1e and Snrk*** (see figure 5.3). This list intentionally ignores the use of up or down regulation, as this is only evidenced by the meta-analysis, whereas the gene expression profiles of the time course experiment may show much more complex patterns depending on time frame. Although many of these genes have been shown to play some role in the IFN-γ signalling cascade or more general immune response to pathogens, the evidence is not conclusive and not specific to the effect of IFN-γ in mouse macrophages, making them a suitable proposition for further hypothesizing and experimental follow-up.

Figure 5.3 Expression profiles of selected new identified ISGs



Gene expression profiles for 15 genes (one of which is represented by two gene probes) which are significant in the consensus result set of meta-analyses, have strongly changing expression profiles in an independent microarray time course of IFN-γ treatment in mouse macrophages (MITCH12), and which are not already known to be interferon stimulated genes (according to Interferome.org). The **X-axis** represents 25 measurement time points from 0 to 720 minutes in 30-minute intervals. The **Y-axis** represents absolute linear-scale

transcription level after data processing and normalisation. Graphs are organised by expression profile pattern: 9 graphs in top section and 2 graphs in the middle section containing genes which are up regulated in meta-analysis results (here visually separated by straightforward up regulation and peaking expression in MITCH12); 5 graphs in bottom section are down regulated in meta-analysis results, although in MITCH12 down regulation is notably transient.

5.6.5 Candidate ISGs based on other evidence

While the previous sections cover the context of meta-analysis findings using selected primary sources, results were also tied into published and on-going research in our group covering immune response to pathogens in immune cells. For these, one can use the results obtained in meta-analyses to emphasize these genes.

A recent paper (Kropp, Robertson et al. 2011) identified interactions between type I and type II immune responses as a contribution to the activation state of macrophages following viral infection. Using IFN-γ treatment and subsequent withdrawal of IFN-γ treatment, this study identified 163 genes that require continuous IFN-γ stimulation in order to maintain their transcription response to viral infection. This number of results would usually be difficult to follow up, but congruency of this set of results with meta-analysis results may help in focusing further investigation. Of the 163 genes identified, 148 are unique genes, and 10 of these show subtle but consistent (that is, identified by meta-analysis alone) gene expression increases in response to IFN-γ treatment. These are ***Klf3, Tox4, Kiaa0247, Rnf14, M6pr, Ahrr, Ankib1, Pnpla2, Stx4, and Hivep2***. Notably, there are no matches against the consensus list of genes down regulated by IFN-γ, which conforms to expectations as the list of 163 genes is only concerned with reversible up regulation of genes.

More in-house research (Blanc 2013, Watterson, Guerriero et al. 2013) suggests links (in response to viral infection) between IFN-γ signalling and sterol biosynthesis, with the former down regulating the latter. This research focuses on the primary sterol biosynthesis pathway consisting of 23 genes. Of these, only two

are also significant in microarray meta-analysis, one each for genes up regulated by IFN-γ (***Ebp***) and genes down regulated by IFN-γ (***Mvd***, a target of Srebp2). While this outcome does not help in further prioritisation of research, it independently verifies it and adds confidence to this analysis and its consistency. New hypotheses on the interaction between IFN-γ and sterol biosynthesis are included in the discussion section of this chapter.

5.7 Master lists of identified candidate genes

As a primary outcome in this thesis, master lists of novel genes identified by microarray meta-analysis and with independent supporting evidence available are included as tables 5.8 and 5.9. A complete list of meta-analysis-only discoveries (consensus lists) *irrespective of supporting evidence* is included in the supplementary material[39].

For tables 5.8 and 5.9, list inclusion criteria apart from significant meta-analysis outcome are: supporting evidence in form of MITCH12 time course expression profiles, relevant literature, STAT1 binding site (GAS motif) in the gene's promoter region, known reversible inhibition of virus replication by IFN-γ. Novel IFN-γ regulated genes are identified as rows without shading (row shading indicates genes are known as stimulated by IFN-γ type I, II, or III). Ordering of genes in theses tables is based on the number of pieces of supporting evidence (equal weight assumed), although this does not imply genes with less supporting evidence or only a particular source of evidence should be discarded from follow-up.

[39] Folder "MetaAnalysisResults", file "CategorisedCumulatedGeneList.xlsx"

Table 5.8 Master list of IFN-γ up regulated meta-analysis-only discoveries with supporting evidence

Gene Symbol	Notable MITCH12 profile	STAT1 binding site	IFNG reversibly inhibits virus	Publications of interest
M6pr	Y		Y	Co-localises with Stat1 and HPIV proteins, allowing virus to block IFN signalling (Schomacker, Hebner et al. 2012)
Rnf14	Y	Y	Y	
Stx4 (Stx4a)	Y		Y	Rate-limiting step of TNFa production in macrophage activation (Pagan, Wylie et al. 2003)
Tbk1	Y	Y		Induction of Ifn-beta, IRF3 through TLR (Hemmi, Takeuchi et al. 2004, Perry, Chow et al. 2004)
Abcc5	Y	Y		
Ahrr	Y		Y	
Batf		Y		Inhibits AP-1 activity and affects T-cell proliferation (Thornton, Zullo et al. 2006). Induced by STAT3 (Senga, Iwamoto et al. 2002). Protein forms complex with IRF4 and JUNB (Tussiwand, Lee et al. 2012)
Cd47	Y			Binds macrophage SIRPa (Oldenborg, Gresham et al. 2001). Listed as ISG (de Veer, Holko et al. 2001)
Clic4	Y			Regulates M0 deactivation (Malik, Jividen et al. 2012). In bacterial M0, induced by LPS and positive regulator of LPS (He, Ma et al. 2011)
Ehd1	Y	Y		
Igbp1	Y			Depletion enhanced induction of STAT1-dependent genes (Fielhaber, Han et al. 2009)
Ikbke	Y			Critical for regulation of IFNb and ISGs (Takeuchi, Hemmi et al. 2004). Restores RIG-I dependent antiviral response on HepC infection (Breiman, Grandvaux et al. 2005). Regulates balance between type I and type II interferon responses (Ng, Friedman et al. 2011).
Klf3		Y	Y	
Lgals8		Y		Inhibition of IL-6 production in M0 (Yang, Jiang et al. 2011)
Marcks		Y		Candidate biomarker for interferon therapy response (Huang, Tu et al. 2008)
Psen1	Y	Y		
Raf1		Y		Component of MAPK cascade (Farrar, Alberolalla et al. 1996, Jesenberger, Procyk et al. 2001)
Stard3	Y			Cholesterol transport (Strauss, Liu et al. 2002)
Tmod3	Y	Y		
Tpm3	Y	Y		
Ankib1			Y	
Arf3		Y		
Atad1	Y			
Capza2	Y			
Casp12				Stimulated by IFN-γ in fibroblasts (Kalai, Lamkanfi et al. 2003)

Fbrs	-	Y		
Fxc1		Y		
Gosr2				In HepG2 cells, TNFa up-regulates Gosr2 (Pandey, Munjal et al. 2010)
Hivep2			Y	
Inpp1	Y			
Itga4				Induced by IFN-γ in monocytes (Waddell, Popper et al. 2010)
Kiaa0247	-		Y	
Ktn1	Y			
Lrp10 (Lrp9)	Y			
Mrpl13		Y		
Ms4a4b	Y			
Obfc2a	-	Y		
Ppap2a	Y			
Slc35b1	Y			
Tox4	-		Y	
Txndc17	-	Y		
Usp12	Y			
Vcan	-	Y		
Bcl3	Y	Y		Protein is known to interact with p50 and p52 subunits of NF-kappaB (Schwarz, Krimpenfort et al. 1997). Suppresses p53 activation and induced apoptosis (Kashatus, Cogswell et al. 2006). Inhibits IL-10 expression in M0 (Riemann, Endres et al. 2005)
Birc2	Y	Y		Up regulation protects B cells from deletion, allows for IgA secretion (Husain, Holodick et al. 2006)
Junb	Y	Y		STAT3 binds to Junb promoter (Oh, Kim et al. 2009)
Pnpla2	Y	Y	Y	
Arl6ip5	Y	Y		
Cflar	Y			Controls NF-kappaB activation in DCs (Golks, Brenner et al. 2006)
Lpp	Y	Y		
Ly6a	Y			T-cell proliferation, regulated by TNFa (Henderson, Kamdar et al. 2002)
Mertk	Y			Regulates TNFa and M0 activation (Camenisch, Koller et al. 1999, Sather, Kenyon et al. 2007).Innate immunity (Behrens, Gadue et al. 2003) Stat1 activation.
Mxd1	Y			Induced by IFN-γ, inhibits M0 mitogenesis (Dey, Kim et al. 1999)
Rgs14	Y	Y		
Sfpi1	Y			Blocks T-Cell development. Regulates macrophage differentiation in response to cytokines. Regulates transcription of IL7Ra (DeKoter, Walsh et al. 1998, DeKoter, Walsh et al. 1998, Anderson, Weiss et al. 2002)
Hnrnph2	-	Y		
Mafk	Y			
Ubr4	-	Y		

Genes in this table are up regulated (statistically significant) by IFN-γ microarray meta-analysis, where this result is also supported by independent evidence. Independent supporting evidence is defined as the gene having a notable expression profile in the MITCH12 microarray time course study (identified through GEViSE tool), containing a STAT1 binding site (GAS motif, identified through oPOSSUM software) in its promoter region, having been identified as reversible inhibitor of viral infection through IFN-γ (Kropp, Robertson et al. 2011), or having had literature related to immune responses -

irrespective of a specific relation to type II response in murine macrophages – published (identified through PubMed with combined searches of the official gene symbol and "interferon"). The number of pieces of supporting evidence are also used to sort genes within the table, with equal weight given to all evidence sources. Gene symbols in **bold** font are those that can be considered high-confidence findings on merit of their identification through all three meta-analysis model families. Shaded rows indicate that a gene is already known as regulated by IFN-γ through the Interferome database (irrespective of it being known as a type I, II, or III regulated gene). A horizontal dash in the MITCH12 column indicates that the gene was not found within the MITCH12 microarray platform, this is different from an unconfirmed expression change ("N")

Table 5.9 Master list of IFN-γ down regulated meta-analysis-only discoveries with supporting evidence

Gene Symbol	Notable MITCH12 profile	STAT1 binding site	IFNG reversibly inhibits virus	Publications of interest
Ptpre	Y	Y		Inhibits JAK-STAT signalling in leukaemia cells (Tanuma, Nakamura et al. 2000). Up-regulated during differentiation and/or activation of macrophages, and its overexpression suppress IL-6 and IL-10-induced JAK-STAT signalling (Tanuma, Shima et al. 2001)
Csnk1e	Y			Control of circadian rhythm through PER2 degradation (Eide, Woolf et al. 2005)
Dbi	Y			Regulated by SREBP-1c in hepatocytes (Neess, Kiilerich et al. 2006)
Galnt6	Y	Y		
Hyal2	Y			Entry receptor for a sheep retrovirus. Disassociation from RON allows activation of Akt and MAPK cascade (Danilkovitch-Miagkova, Duh et al. 2003)
Lta4h	Y			Proinflammatory lipid mediator (Billharz, Zeng et al. 2009)
Map2k3		Y		Essential for TNF-stimulated p38 MAPK activation (Brancho, Tanaka et al. 2003). Related to IL-12 activated immune response pathway through p38 and STAT4 (Watford, Hissong et al. 2004)
Map4k4		Y		Activation of TNFa promoter (Mack, Von Goetz et al. 2005)
Nfatc1		Y		May express IL-4 independent of Stat6 (Wang, Kusam et al. 2006)
Nfya		Y		Reduces induction of JunB by IL6 (Sjin, Krishnaraju et al. 2002)
Nisch	Y	Y		
Pde7a		Y		In T-cells, increase correlates with IL2 (Li, Yee et al. 1999)
Pgrmc1		Y		Putative steroid-binding protein (Nolte, Jeckel et al. 2000). Binds to SCAP and Insig-1 (both part of a complex in sterol

				biosynthesis) (Suchanek, Radzikowska et al. 2005)
Sirpa	Y			Cd47 binds macrophage Sirpa protein and suppresses phagocytosis (Oldenborg, Gresham et al. 2001)
Vamp3		Y		Related to macrophage phagocytosis (Allen, Yang et al. 2002)
Arid2		Y		
Asxl1		Y		
Atp6V0a1	Y			
Cdt1	Y			
Dnajc11		Y		
Dock7		Y		
H1fx		Y		
Isca1	-	Y		
Lrrfip1				Repressor of TNFa (Suriano, Sanford et al. 2005). siRNA against LRRFIP1 was found to inhibit type I IFN production induced by [listeria] (Rathinam and Fitzgerald 2011, Thompson, Kaminski et al. 2011)
Mapk1				Transcriptional repressor of interferon (gamma) signalling (Hu, Xie et al. 2009)
Mvd	Y			
Nudt5		Y		
Pik3r2	Y			
Ppan	Y			
Prpf19		Y		
Ptp4a1		Y		
Reep6		Y		
Rrp1b	-	Y		
Snrk	Y			
Snx21	-	Y		
Synj2	Y			
Tspan5	Y	Y		
Usp2		Y		
Usp48		Y		
Wdr48	Y			
Capn2		Y		
Degs1	Y			
Elk3	Y			(note: strong MITCH12 expression profile, but up-regulated in time-course while down-regulated in meta-analysis)
Gps1		Y		
Hey1		Y		
Rab11a				Promotes phagocytosis in M0 (Cox, Lee et al. 2000)
Tshb		Y		

Genes in this table are down regulated (statistically significant) by IFN-γ microarray meta-analysis, where this result is also supported by independent evidence. Independent supporting evidence is defined as the gene having a notable expression profile in the MITCH12 microarray time course study (identified through GEViSE tool), containing a STAT1 binding site (GAS motif, identified through oPOSSUM software) in its promoter region, having been identified as reversible inhibitor of viral infection through IFN-γ (Kropp, Robertson et al. 2011), or having had literature related to immune responses - irrespective of a specific relation to type II response in murine macrophages – published (identified through PubMed with combined searches of the official gene symbol and "interferon"). The number of pieces of supporting evidence are also used to sort genes within the table, with equal weight given to all evidence sources. Gene symbols in **bold** font are those that can be considered high-confidence findings on merit of their identification through all three meta-analysis model families. Shaded rows indicate that a gene is already known as regulated by IFN-γ through the Interferome database

(irrespective of it being known as a type I, II, or III regulated gene). A horizontal dash in the MITCH12 column indicates that the gene was not found within the MITCH12 microarray platform, this is different from an unconfirmed expression change ("N")

5.8 Discussion

This chapter aims to answer the primary question if meta-analysis of a modestly-sized set of similarly-themed small microarray transcription studies can reveal new biological research spotlights that cannot be obtained by the analysis of the individual studies alone. Before discussing outcomes related to specific genes, a number of issues relating to the information and interpretation methods in this chapter require consideration.

5.8.1 Discussion of issues relevant to result interpretation

Complete type II interferon response. An alternative and worthwhile pursuit for a more biological thesis would consist of a complete analysis of all genes regulated directly and indirectly by IFN-γ. This would be accomplished by considering all genes with a statistically significant meta-analysis result, irrespective of a gene also being identified as statistically significant in the analysis of any one of the microarray studies. However, in this thesis the emphasis must lie on the statistical performance of meta-analysis beyond the prospects of single-study analyses. That is, what biology does a meta-analysis uncover that cannot be uncovered by an individual study? Given this focus, the results in this chapter cannot cover the complete transcriptional response to IFN-γ and will lack well-known actors or modifiers of the type II response, because these will almost by definition be those genes showing strong transcriptional responses to IFN-γ stimulus, and therefore be among the first discoveries made in any microarray study (or other platform). This affects reference comparisons, where gene list overlap with known ISGs would likely be much higher if individual study results were included. It also affects summary output in form of STRING (Franceschini, Szklarczyk et al. 2013) protein interactions within the identified gene lists, with many potential interactants

missing. This is one of the major considerations to be aware of in any reading of this chapter.

Biological framework. All microarray studies used in meta-analysis are investigating the effect of IFN-γ (with different doses and time points) on murine macrophages. This is a more specific framework than the research into interferon stimulated genes (ISGs) here used as reference in form of the interferome database and in-silico STAT1 binding site identification, because the latter also covers type I and type III responses and distinctions are not always clear. This is not necessarily a crucial factor to be taken into account, as many interactions between these signalling pathways continue to be found. Because interferons of all types may have an effect on or be influenced by a type I interferon response, it is here not considered useful to remove findings not obviously related to type II interferon. A similar argument applies to accepting reference findings from cell types other than macrophages, because aspects of JAK-STAT or other signalling found in immune cells other than macrophages may also apply to macrophages but lack current confirmation, encouraging further laboratory testing of hypotheses.

Interferome database. Although probably the largest resource of ISGs, this database cannot be considered comprehensive. While it may have considerable overlap with ISGs other research groups have arrived at through proteomic, single-gene or other laboratory platforms, there is little doubt that numerous important candidates are not included. This does lessen the interpretation of "candidate ISGs" used in this chapter, in that a gene may not be contained in the interferome database and on cursory investigation not known to a given researcher, but it may be an obvious and known ISG to others, based on published research they are familiar with.

Objective standards. The statistical meta-analysis of microarray studies can be considered objective, but there are two areas in the biological interpretation of their results that can suffer from subjectivity. One is the use of temporal gene expression profiles as provided by MITCH12 (see 5.4.1). Measured across 12

hours, few genes exhibit entirely straightforward expression level changes. The interpretation of those is therefore dependent on an individuals' perception of what constitutes a clear rise, drop, peak or trough, or what constitutes an overall profile of interest. In order to alleviate this, the results discussed in more detail in this chapter are limited to expression changes that are above 2-fold (up or down). Summary tables will also contain information of lower confidence. The second area is gene annotation, where a compromise must be found between thoroughness of investigation and inclusiveness of results. Because of the number of genes involved, not all available sources of annotation or literature can be assessed for each gene. Information provided here is therefore based on a subjective assessment of relevance. This includes annotation for genes obtained for the human rather than mouse genome, implying homologous function which is not necessarily true.

Gene sets. A by-product of comparing multiple meta-analysis models is the number of gene lists this produces. While it is possible to broadly assess their differences and features (chapter 3), it makes a thorough look at biological context of each list difficult. The gene list consolidation into one list of 106 IFN-γ up regulated and 152 down regulated genes was therefore necessary but will of course not allow further conclusions on which meta-analysis provides the most biologically relevant results. In that context, an interesting but here omitted list (included in supplementary material) are those genes only identified by a single meta-analysis model to the exclusion of all others.

Transcriptomics. A general complication of "omics" approaches is the discrepancy between hypothesis driven and data driven research. Microarray studies are not initialised with one or just a few hypotheses regarding the role of particular genes in a system in question. This thesis covers the analysis of such data and will therefore always provide data driven (potential) answers to questions that were not posed. Apart from the use of highly domain-specific knowledge allowing immediate identification of interesting (where this entirely depends on the operator) genes amongst results, the optimum achievable results are usually gene

lists with varying amounts of prioritisation. It is always the case that these require experimental follow-up in prospective laboratory studies, and this thesis is no exception.

Overlap between meta-analysis and known biology. The biological context of microarray meta-analysis results appears to be relevant to the biological question of transcriptional responses to IFN-γ stimulation of cells. This statement is supported by non-random overlap between differentially regulated genes obtained through meta-analysis and genes already known to be part of IFN-γ signalling and transcription. This point is strengthened by the fact that gene lists obtained through meta-analysis do (by design) not contain any of the statistically significant results that are discovered in any individual microarray studies. One side effect of this is ineffectiveness of gene set enrichment tests. While tools such as DAVID[40] are often used to identify significant clusters of functional annotation categories with submitted gene lists, this is here problematic because meta-analysis consensus lists (with 106 up and 152 down IFN-γ down regulated genes) do not contain any of the genes statistically discovered in individual studies and are therefore "de-enriched" for direct IFN-γ transcriptional responses. Additionally, with these consensus lists expected to produce IFN-γ stimulated genes as yet unconfirmed by other research, existing gene annotation categories may not reflect their function within type II responses. Use of DAVID in this case results in a majority of genes without current annotation classes, with the remainder of genes belonging to significant clusters of important but unspecific nature, for interferon up regulated genes these are transcription regulation, intracellular transport, protein localisation, membrane, transmembrane and nucleotide binding, for interferon down regulation these are phosphorylation, kinase activity and membrane.

5.8.2 Discussion of genes in biological context

Interactions between interferon signalling and sterol biosynthesis. Several of the potential new ISGs suggested by meta-analysis correspond to a recently suggested

[40] http://david.abcc.ncifcrf.gov/

role of interferon as a regulator of antivirally acting 25-hydroxycholesterol production in macrophages (Blanc 2013). *Mvd* and *Ebp* are already verified interactors in sterol biosynthesis and meta-analysis here confirms opposite but significant reactions to IFN-γ (with Mvd down regulated and Ebp up regulated). However, meta-analysis also provides novel genes for interactions between type II interferon signalling and sterol biosynthesis. *Stard3* is identified as up regulated by IFN-γ, which is strongly confirmed in the MITCH12 time course study. Literature suggests a role for *Stard3* in cholesterol transport (Strauss, Liu et al. 2002) and an only unimportant role in sterol biosynthesis (Kishida, Kostetskii et al. 2004). In macrophages *Stard3* overexpression does not affect SREBP2 transcription but represses lipogenesis (Borthwick, Allen et al. 2010). However, these studies do not cover interferon signalling in response to infection, making this a worthwhile hypothesis for further laboratory work.

Pgrmc1 (here down regulated by IFN-γ and tendentially but not conclusively confirmed by the time course study) has been shown to directly interact with SCAP and Insig-1 (Suchanek, Radzikowska et al. 2005) and to promote sterol biosynthesis in embryonic kidney cells (Hughes, Powell et al. 2007). If these roles can be confirmed in macrophages, a new hypothesis would consist of IFN-γ down regulating *Pgrmc1*, and therefore affecting the SCAP, Insig-1, SREBP complex in its regulation of sterol synthesis.

Two additional new hypotheses linking IFN-γ signalling with sterol biosynthesis could be based around the expression pattern similarity of both *Itga4* and *Obfc2a* to *Ch25h* (STAT1 induced, couples interferon response to 25-hydroxycholesterol) when analysed by a clustering algorithm in a microarray time course study (not the MITCH12 study used for validation in this chapter) involving 8 hours of treatment with IFN-γ. This information is obtained from table S1 and figure 6 of Blanc et al (Blanc 2013) but is not verifiable through the MITCH12 time course and as such a weaker proposition.

Lrp10 (alias *Lrp9*) as a membrane protein may serve in the uptake of apoE-containing lipoproteins (Sugiyama, Kumagai et al. 2000) and SREBP2 protein mediates *Ldlr* gene transcription in hepatic cells (Huang, Zhou et al. 2008) and in humans Ldlr protein has been shown (Brodeur, Theriault et al. 2012) to have a

subfamily containing Lrp10 (Lrp9 in mice). While this is far from conclusive, strong up regulation by IFN-γ is confirmed in MITCH12, pointing to a possible hypothesis around Lrp10 taking part in the interaction between interferon signalling and sterol biosynthesis over SREBP2.

Three further meta-analysis outcomes related to recent research into sterol biosynthesis are *Galnt6*, *Rab11a* and *Golga4*, all down regulated by IFN-γ (Golga4 is not part of the master table due to lack of supportive evidence in the used references) but awaiting further research.

Another IFN-γ induced (as identified by meta-analysis and verified by MITCH12 time course data) gene is *Ehd1*, its protein has been suggested to associate with fibroblast LDL receptor and regulate cholesterol uptake (Naslavsky, Rahajeng et al. 2007) and be involved in the transport of Rab11 containing vesicles (Naslavsky, Rahajeng et al. 2006), which would have relevance to the above mentioned *Rab11a* in our in-house research on crosstalk between type II interferon immune response and sterol, especially since there appears to be no published research on *Ehd1* in context of cytokine activation available at this time.

Crosstalk between type I and type II interferon signalling. Meta-analysis here confirms potential mechanisms of crosstalk between these two pathways as described in Kropp et al (Kropp, Robertson et al. 2011). Induction of a known type I interferon signalling gene - *Tbk1* – identified in that study under low IFN-γ dose is here confirmed by meta-analysis of several studies. Although the meta-analysis includes the microarray study used in that paper, it is not statistically identified in that study (due to more stringent criteria). However, the effect of IFN-γ on *Tbk1* is consistent across microarray studies (ranging from low dose to very high dose of IFN-γ) and therefore identified through meta-analysis.

Not an outcome related to the Kropp et al study, *Ikbke* (alias *IKKε or IKK-i*) is a known regulator of antiviral signalling and has recently been proposed (Ng, Friedman et al. 2011) as a regulator of the crosstalk between type I and type II interferon pathways in that by phosphorylating STAT1 it suppresses formation of STAT1 homodimers and thus shifts the transcriptomic response to type I interferon. Meta-analysis would here underline the fact that in absence of viral

infection, treatment with IFN-γ does lead to small but consistent up regulation of *Ikbke* where the question would be if this is to regulate type I interferon stimulated genes and/or to act as feedback control for type II interferon stimulated genes. *Ikbke* is also suggested to be functionally equivalent (production of interferon beta through IRF3 or IRF7) to *Tbk1 (Ikeda, Hecker et al. 2007)*. The presence of both *Tbk1* and *Ikbke* in meta-analysis-only results implies that in many transcriptional studies they will not be identified as statistically significant, which is reinforced by their absence from the (microarray-based) Interferome database. This would support a closer investigation into their combined effect on type I- type II crosstalk.

Batf. It has been shown that IFN-γ suppresses IL10 production by suppressing AP-1, thus augmenting TNFa induction (Hu, Paik et al. 2006). With BATF known as an inhibitor of AP-1 mediated gene expression (Deppmann, Thornton et al. 2003), consistent up regulation by IFN-γ in microarray meta-analysis may suggest that this regulation is involved in crosstalk between type II interferon and IL10 signalling pathways.

Clic4 is known to be up regulated through LPS or combined LPS and IFN-γ treatment (Malik, Jividen et al. 2012), although its function is still under investigation. He et al (He, Ma et al. 2011) investigated its role in infection and found the *Clic4* gene to be induced by LPS and to act as a positive regulator of LPS signalling through IRF3 phosphorylation. The induction of *Clic4* in macrophages is here confirmed by microarray meta-analysis (and reinforced by the MITCH12 time course), using a range of treatment times and doses. Notably, none of the microarray studies use LPS in their treatment regime, suggesting that it is not only LPS signalling through TLR receptors that is involved, but that type II interferon alone may have downstream effects on IRF3 through *Clic4*.

In meta-analysis, *Ptpre* is found to be down regulated by IFN-γ, in the MITCH12 time course this is also true for the first six hours, after which it returns to original levels. In published research, a cytosolic isoform of its protein has been found to

inhibit IL-6 and IL-10 induced JAK-STAT signalling (Tanuma, Shima et al. 2001) by preventing phosphorylation of STAT3. The same research also suggested that it does not affect STAT phosphorylation by type I or type II interferon signalling. While this research was not in macrophages, it may allow to hypothesize that IFN-γ down regulation of *Ptpre* could aid in the induction (or prevention of inhibition) of JAK-STAT signalling, and based on MITCH12 data, that this effect is transient.

Csnk1e is a known circadian clock gene (PER2 associated), and limited research into the interactions between cytokine signalling and clock genes has shown that TNFa and IL-1b suppress clock genes (including *Per2*) in mouse fibroblasts (Cavadini, Petrzilka et al. 2007); that murine macrophages contain an intrinsic clock that may regulate innate immune responses (Hayashi, Shimba et al. 2007); and that antiviral responses could be under circadian control via the JAK-STAT or NF-kB pathway in human T-cells (Bollinger, Leutz et al. 2011). It is currently not clear how extracellular IFN-γ suppresses circadian clock associated gene *Csnk1e* but the outcome from this study would suggest that it contributes to the antiviral state through immediate-early mechanisms, with MITCH12 showing that it starts returning to original expression levels after 4 to 5 hours after IFN-γ treatment.

Down regulation. While the initial set of genes that meta-analysis identifies as IFN-γ down regulated is larger than the set of up regulated genes, the amount of available corresponding evidence (or lack thereof) is less and may be related to common research biases favouring up regulation as reported outcomes. While this makes prioritisation of research spotlights problematic in general, confirmation for down regulation is still provided by the independent time course microarray study, possibly raising the importance of also following through research on IFN-γ down regulated genes. It is also notable that all of the genes identified as down regulated by meta-analysis and strongly confirmed by MITCH12 time course have a transient down regulation and reverse to their original expression levels by 12 hours after treatment (figure 5.3). It is unclear if this is a chance finding, if it is related to the nature of observable down regulation in microarray studies; or indeed if this is biologically meaningful.

JAK-STAT. Two unexpected results in terms of their presence in the meta-analysis-only results are *Sfpi1* (a known regulator of STAT) and *Junb* (known to be induced by STAT1 or STAT3). In theory, these should have been identified as statistically significant in some of the individual microarray studies as well as meta-analysis and therefore not be present in these results. This indicates inconsistent responses to IFN-γ treatment across microarray studies, possibly brought about by different treatment regimes. It is also possible that the use of different measurement time points results in those probes as sometimes up, sometimes down regulated, for which there is a slight indication in the microarray time course data (where expression levels fall and rise again over time).

Conclusion. Many priority (higher up in the master tables) genes identified would appear to fit directly into interferon-related processes or into processes which have recently been identified within the Ghazal lab or elsewhere to be relevant to the interferon response (sterol pathway, type I interferons). With meta-analysis identifying genes of lower transcriptional activity than those in the main IFN-γ response (which were subtracted from these results prior to interpretation), the mechanism by which individual genes fit into the established JAK-STAT pathway and set of ISGs is not easy to identify. If one assumes that meta-analysis was successful in identifying novel gene functions, then their ultimate confirmation within the system of interest (type II immune response in murine macrophages) relies on prospective laboratory studies, with existing research on these genes often only confirming relevance if allowing for entirely different treatments, cell types or organisms. This chapter demonstrates that a statistical meta-analysis of even a small and heterogeneous set of microarray studies can provide results that match the expected biological framework and provide novel hypothesis for experimental validation. However, statistical meta-analysis results are not self-evident or free of false positives, making a secondary review of results with a biological emphasis a necessity. This requires analysts to either obtain a sufficient amount of specialist biological knowledge or to collaborate very closely with biologists in order to avoid unexploited gene lists.

Chapter 6

Network analysis of an IFN-γ time course study

This chapter test the ability of an enhanced network analysis to identify genes or gene sets with coordinated activity over time, and develops a hypothesis for *Trafd1* involvement in sterol biosynthesis upon IFN-γ activation of macrophages.

A network analysis of IFN-γ microarray time course data is complementary to the case-control nature of microarray studies used in meta-analysis. Where meta-analysis in this thesis can only combine snapshots (with individual study design prescribing different doses and different measurement time points) of the effect of IFN-γ treatment on macrophages, a time course allows identification of potential causal relationships between genes by measuring their co-expression levels over a period of time. While lacking the statistical rigour of meta-analysis, network analysis provides a nuanced data set for validation and discovery of new biology.

The work in this chapter is motivated by a) the availability of time course microarray data with a relatively large number of time points, b) the use of the same data (although only the IFN-γ time course) for direct validation of meta-analysis results (see chapter 5), and finally c) the supposition that a methodical (although not objective or statistically rigorous) network analysis can identify meaningful gene set expression patterns in a time course study. An attempt is made to reduce common network (and clustering) analysis bias in form of guiding the analysis towards expected biological results, by avoiding trial-and-error parameter choices that are judged on expected biological outcomes. Instead, parameters are chosen based on the computational ability of the applied processing steps. This chapter does not propose a correct or optimal set of parameters

applicable to other data sets, but a general multi-step approach that in this case leads to interesting high-level results.

Gene expression profiles obtained from this microarray time course study into the time-dependent effects of IFN-γ or murine cytomegalovirus were also used as part of the independent validation of meta-analysis results (chapter 5), but in a purely gene-by-gene approach. This chapter expands on the results by utilising similarity-structures across genes in a time course (network analysis).

6.1 Macrophage activation and function

Established models for classical and alternative macrophage activation are introduced in chapter 1, and here supplemented by further pathway information relevant to gene transcription network analysis.

The activation and function of macrophages is a complex process, centred on intricately coordinated networks of gene and protein expression that balance the level and duration of anti-inflammatory and pro-inflammatory macrophage functions in order to maximise harmful effects towards pathogens while minimising harmful effects towards the host. Many immune cells can produce multiple cytokines and chemokines, and cytokines and chemokines can be produced by multiple types of cells. Many cells will produce proteins that can both boost and reduce further immune responses (such as type I IFN). Classical activation of macrophages includes not just cytokine-mediated activation (primarily type II IFN signalling), but also activation through pathogen recognition (TLR signalling) and activation through immune receptors of other immune cells (ITAM signalling). While many aspects of signalling cascades and downstream products of individual signalling pathways are understood, less is known about the complex crosstalk between them (Hu, Chen et al. 2007). Feedback loops, crosstalk and overlapping cell functions result in highly complex scenarios driven by many factors (time, expression levels, location). The description provided below can therefore only serve as a simplifying overview.

Macrophages can be activated by recognition of host factors (cytokines, primarily IFN-γ), recognition of pathogens by TLRs or recognition of other immune receptors (e.g. present on B or T cells). Respectively, these invoke IFN signalling, TLR signalling and ITAM signalling in the macrophage.

Member genes and proteins of signalling pathways differ and the downstream effect of macrophage activation depends on many factors, but it can be summarised as production of cytokines (TNFα, IFNα, IFNβ, IL-1b, IL-6, IL-10, IL-12, IL-18), cell surface molecules (MHC), growth factors (TGFβ), and other molecules (nitric oxide, reactive oxygen). These four groups affect, respectively, inflammatory processes, antigen presentation, tissue repair and killing of pathogens. While this grouping may suggest mutual exclusivity of function, this is not necessarily the case (e.g. nitric oxide can activate transcription factor NF-kB).

Products of activated macrophages drive both pro-inflammatory and anti-inflammatory processes, where the proportion will depend on the stage or level of infection, with the aim of mounting a sufficiently sized response to pathogens before returning cytokines to homeostatic levels after pathogen eradication and tissue repair.

This fine-tuning of increasing or decreasing inflammatory processes depends on positive and negative feedback loops and interactions between TLR, ITAM and IFN signalling, and the balance between macrophage-activating and macrophage-suppressing factors. A prominent example of the latter is the balance between IFN-γ and IL-10, with IFN-γ inhibiting IL-10 production and driving expression of pro-inflammatory genes via activation of STAT1, and with IL-10 suppressing TLR signalling and driving expression of anti-inflammatory genes via STAT3.

Gene co-expression analysis of a microarray time course study in theory enables global identification of coordinated gene expression as it occurs where a) macrophages are immune-activated without presence of a pathogen (treatment with IFN-γ), b) macrophages are activated by pattern recognition receptors through viral infection in the absence of (initial) IFN-γ and c) IFN-γ activated macrophages are challenged with the same viral infection. Gene co-expression measured for these three conditions would be expected to identify different

patterns of regulation, with condition (a) initially only engaging type II signalling, but viral recognition (b) engaging TLR signalling. The study described here is referred to as MITCH12 and described below.

6.2 Microarray time course study

MITCH12 is a microarray study consisting of three independent time courses, which measure the effect of IFN-γ, the effect of murine cytomegalovirus (MCMV), and the combined effect of IFN-γ and MCMV. For each time course, expression level measurements are taken over 12 hours, in 30-minute intervals and starting at 0 hours (i.e. 25 time points per time course). The study is described in detail and with a graphical overview in chapter 2, section 2.8.2.

Classical macrophage activation is achieved when both an IFN-γ signal and pathogen recognition (LPS signal) occur, which in case of MITCH12 is provided only by the third condition. However, even this experiment condition does not establish full macrophage function, due to completely missing interactions with other types of immune cells. Therefore, this study's aims are in the identification of isolated signalling pathways and limited interactions between them in the absence of full system complexity.

6.3 Network analysis introduction

It is common in microarray studies to identify genes that behave similarly across samples (or samples that behave similarly across genes) in terms of their measured expression level. When using gene (or sample) expression profiles to determine relatedness of genes, network analysis has the goal of identifying co-expressed sets of genes (or samples). Unlike related methods (clustering algorithms like K-Means or Support Vector Machines), network analysis accepts more diverse inputs, such as defining relationships between genes in other ways than expression profile distance/similarity (e.g. known interactions between their proteins). Partitioning into gene sets is performed by topographical rather than clustering

algorithms, that is, genes are visually arranged in a 2D or 3D view based on their measure of relatedness, genes are represented as single nodes rather than gene expression profiles, and the strength or nature of relatedness between genes is indicated as lines (edges) between those nodes.

In this chapter, visual co-expression networks are identified on the basis of gene-gene correlation. The entirety of all gene-gene Pearson correlations is referred to as adjacency matrix or correlation matrix, with the resulting network often referred to as a "guilt-by-association" network. The adjacency matrix is here enhanced to remove redundant gene-gene relationships (Partial Correlation & Information Theory, described in detail below) and the gene networks visually laid out through a topographical algorithm contained in Cytoscape (see discussion). The use of simple adjacency matrices (characterised by measuring similarity rather than distance between genes) is externally motivated in that Pearson correlation (as an R analysis function) has previously been parallelised (coded for use on large computing systems with multiple CPUs) in collaboration[41] between DPM and the Edinburgh Parallel Computing Centre, with the purpose of facilitating computation of large adjacency (or distance) matrices as they can occur with high-throughput biological platforms.

6.4 Network analysis

6.4.1 Input data

The input microarray data set was processed (background corrected, normalised and $\log_2$ transformed) as described in chapter 2, and performed by another member of DPM for a previous analysis (Mizanur Khondoker, unpublished work). This also included a non-specific filtering step to reduce the number of gene probes from the original full set of 22393 gene probes to 14299 gene probes with "non-flat" expression profiles (also excluding control probes).

[41] www.r-sprint.org

6.4.2 Computation of adjacency matrices

For each microarray time course, Pearson's product-moment correlation coefficient *r* is computed for all possible gene-gene pairs (14819 x 14819 = 219,602,761) across 25 time points. This identifies positive or negative linear relationships between gene expression profiles or vectors, with *r*= 1 signifying identical profiles and r=-1 signifying exactly opposite profiles. It should be noted here that the use of correlation coefficients means the absolute expression level of a gene is irrelevant and it is only the "shape" of the expression profile that determines the correlation coefficient.

6.4.3 Size reduction of adjacency matrices

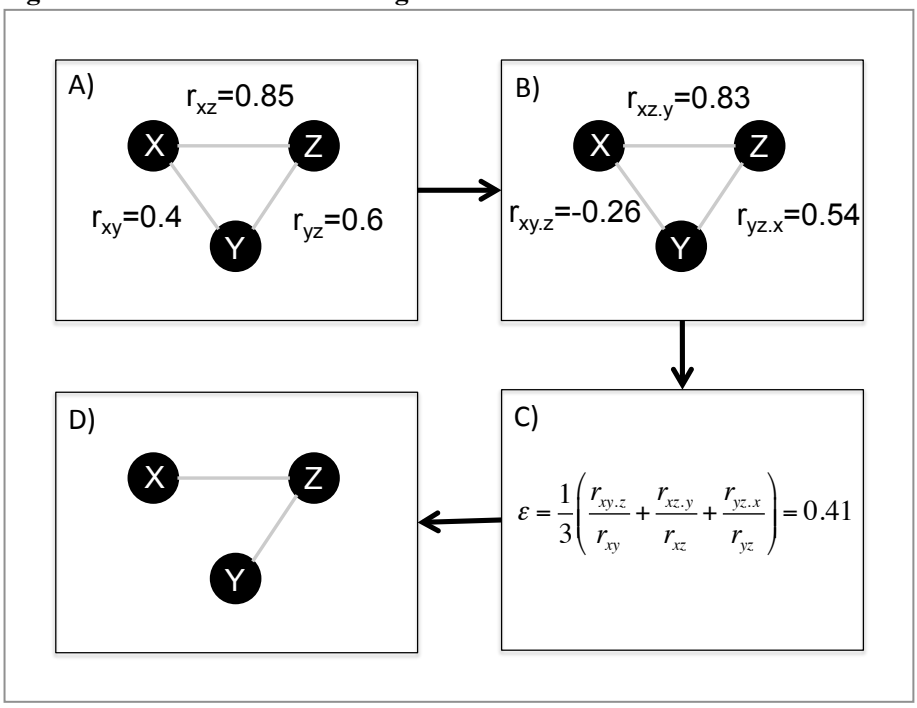
An arbitrary but necessary (for computation efficiency of the following and other) step is introduced to remove genes that do not correlate well with any other gene. In order to remain in the adjacency matrix, a gene has to satisfy a minimum requirement of $|r|>=0.95$ for correlation with at least one other gene.

6.4.4 Removal of redundant gene-gene correlations through PCIT

At this stage, the gene adjacency matrix contains all (filtered) gene-gene Pearson correlation coefficients, but the absence of any weighting means that if one were to draw a network of all these genes now, it would contain both essential and redundant information. "Redundant" here amounts to a lack of causal inference regarding the correlation of any two genes, or technically, the association between two genes depends on a common third (correlated to both of those genes). For example, if gene X is correlated with gene Y, and a gene Z is correlated with both X and Y, the question arises if the correlation between X and Y is causal in itself or if the real causal relationship is between X and Z and Y and Z. Rather than imposing further arbitrary correlation thresholds to separate "good" from "bad" correlations and increase direct causal relationships in the data, approaches have been developed that aim to empirically determine this distinction from the data and therefore lead to better (but not perfect) causal relationship content in the

resulting gene network. This is here done using PCIT, a partial correlation and information theory algorithm (Reverter and Chan 2008). Briefly, for each trio of genes, this calculates the original 3 Pearson correlation coefficients between them, in addition to the 3 partial correlation coefficients (correlation between two genes with removal of the effect exerted by a third gene). The average ratio between original (0^{th} order) correlation coefficients and partial (1^{st} order) correlation coefficients is then used to inform a threshold based on which an association between the two genes in questions is judged significant or not. This is repeated for each possible gene trio, meaning that each gene-gene pair is tested for the influence of any other individual genes in the data set. An example of this is provided in figure 6.1.

Figure 6.1 Overview of PCIT algorithm



PCIT combines partial correlations with a information-theory approach in order to remove indirect gene-gene relationships. A) For any trio of genes (this example uses genes X,Y,Z) in the microarray data set, the three possible Pearson correlation coefficients are computed, denoted as r_{xy}, r_{xz} and r_{yz}. They contain no information about the linear association between two genes with respect to the contribution of a third gene to this association. B) Partial correlation coefficients are computed for the same gene-gene associations, taking into account their possible dependence on a third gene. The new coefficients are denoted as $r_{xy.z}$, $r_{xz.y}$ and $r_{yz.x}$,where the last letter indicates the gene used for conditioning. C) Ratios between the partial correlation coefficient and full correlation coefficient are computed for each gene pair and subsequently averaged as "tolerance level" ε. D) For each gene trio, the decision to keep or discard a link between two genes is

based on the original correlation exceeding ε in the other two gene pairs, in this case leading to removal of the edge between X and Y. The process is repeated for all possible gene trios in the data set.

6.4.5 Basic visualisation of all remaining gene-gene associations

Before assessing visual networks based on remaining correlation coefficients between gene pairs, general correlation structure is assessed through clustered heatmaps of the adjacency matrix.

6.4.6 "Guilt-by-association" networks

Adjacency matrices for each time course were reformatted for import into Cytoscape, with correlations serving as edge weights in 2D networks. Apart from colour-labelling and graph rotation, all node positioning was entirely left to the topographical algorithm, that is, no further manual manipulations (e.g. individual node repositioning) were applied to the algorithm-generated network layout.

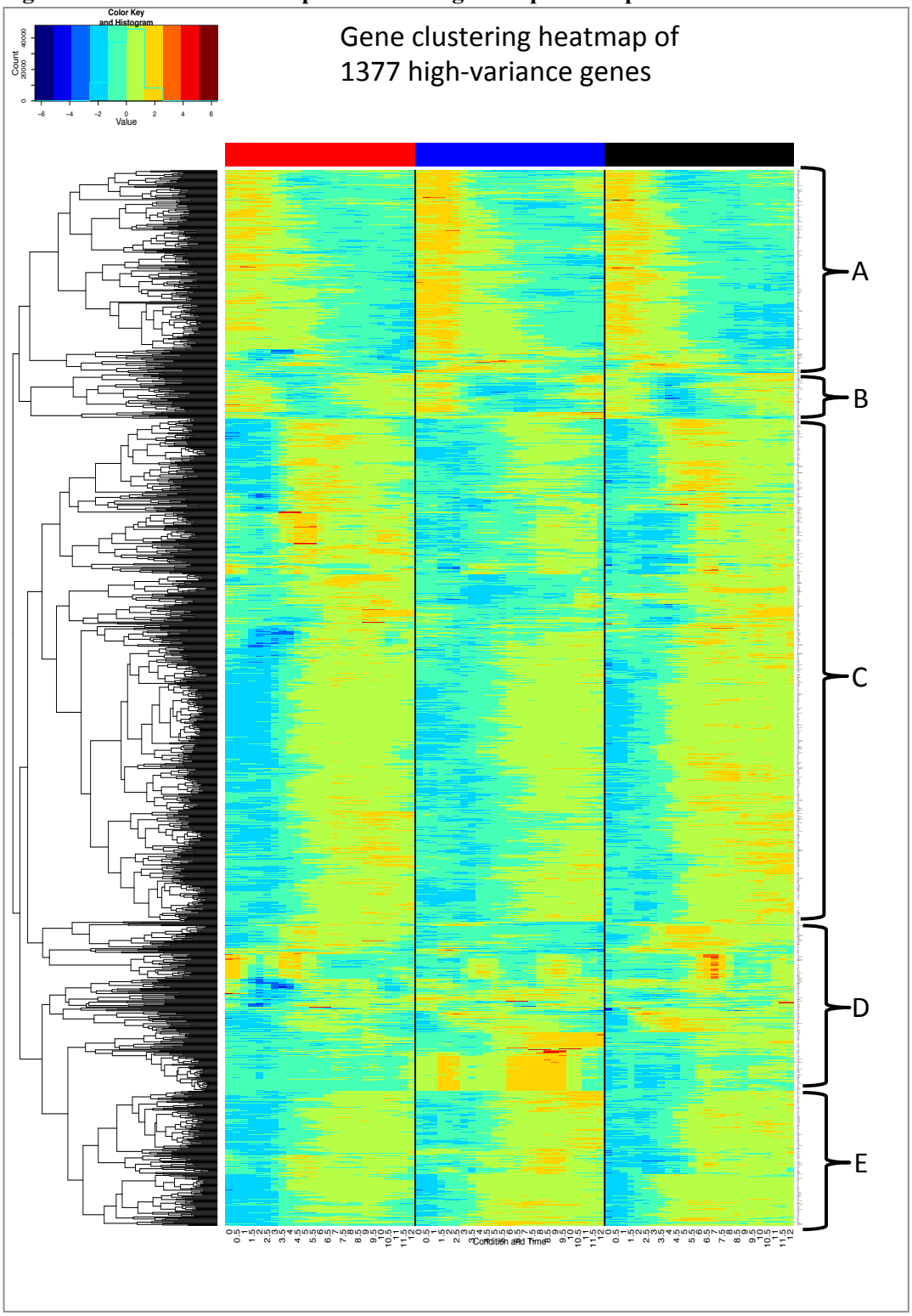
Code for processing of data prior to network analysis can be found in R script "Analysis.R" in the supplementary electronic material.

6.5 Results

Before constructing a co-expression network, global structures in the three time courses are visualised through a heatmap showing clustered scaled gene expression levels. The full data set cannot be used for this purpose because a) computational capacity to perform clustering is usually limited to ~2000 genes (on the computer used for analysis) and b) noise in form of genes never changing expression levels (low variation) or never exceeding some expression threshold (low intensity) obstructs clustering algorithms. Here, the 2000 most variable genes (standard deviation, measured across all three time courses) are further reduced to exclude genes without a single expression level measurement above the maximum

of all 75 sample medians, resulting in 1377 gene probes for clustering shown in figure 6.2.

Figure 6.2 Overview heatmap of clustered gene expression profiles

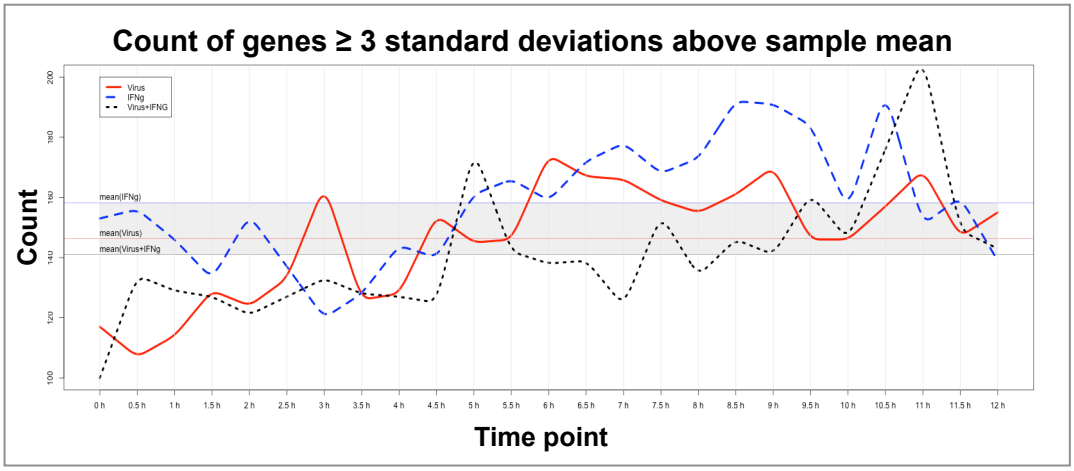


Agglomerative hierarchical clustering of 1377 genes probes with high variation and above-median intensity in at least one sample. Genes in rows, samples in columns. Time courses are identified by the colour bar on top of the graph (red=virus, blue=IFN-γ, black=virus+IFN-γ) and arranged in chronological order from 0h on the left to 12h on the right. Genes expression values are scaled (across all 75 samples, mean=0, standard deviation=1), the resulting z-score is represented by colours: red cells= expression above mean for this gene, blue cells= expression below mean for this gene. The added annotation on the right indicates subjective gene clusters of interest, with A= genes down regulated over time, B=genes with transient down regulation, C= up regulated genes, D= different regulation patterns between time courses, and E=C.

Figure 6.2 is here added to provide a general data overview. While there are clusters of interest amongst the most variable genes, such as distinct sets of up (C,E) and down regulated genes (A,B) and in particular sets of genes with different expression patterns in the three conditions (D), this is part of a general analysis and not part of the network generation as discussed in this chapter.

The general overview can further be elaborated on by finding indications for how the three treatments differ, and a summary graph of genes expressed above a given threshold at each time point and for each condition is presented in figure 6.3. Based on the evidence from highly-expressed genes, gene activation is time-dependent and all three conditions are reasonably similar (within about 50 genes difference at all time points). In terms of a conventional analysis looking at high expression and high expression changes, this would conceivable limit analytical results to a relatively small set of genes. This serves as partial motivation to investigate if the use of gene correlation patterns can provide more information.

Figure 6.3 Count of highly-expressed genes per time-point



For each sample (= time point), z-scores were calculated for each gene probe. Highly expressed genes were defined as those for which expression levels are greater than or

equal to three standard deviations above the sample mean (z>=3). X-axis is divided into 25 half-hourly time points, Y-axis is the count of highly expressed genes at that time point (the scale shown ranges from 100 to 200). Red line represents the virus time course, blue line the IFN-γ time course and dotted black line the combined virus+IFN-γ time course. Horizontal reference lines are the time course average counts across all 25 time points (top=IFN-γ, middle=virus, bottom=virus+IFN-γ).

6.5.1 Size reduction of initial adjacency matrices has notably different effects on the IFN-γ time course

Initial filtering of genes reduces the starting set of 14299 genes in the three time courses to 1779 (IFN-γ), 2834 (MCMV) and 2911 genes (combined interferon treatment and MCMV), respectively. This simple filtering out of genes without at least a single correlation to another gene better than r=|0.95| reduces the computational burden from 204 million gene pair correlations to 3.2, 8.5, and 8 million, respectively. It should be noted that adjacency matrix of course contain self-to-self and duplicated (reversed) gene-gene correlation pairs, meaning the above sizes are n^2 (where n is the number of individual genes in data set), but from a network point of view only $\frac{n^2-n}{2}$ are unique and required. However, the full number of computations is still necessary due to the following processing step requiring a complete correlation matrix as input. Subsequent removal of redundant gene-gene correlations identified by PCIT further reduces the size of adjacency matrices (which is the number of gene-gene correlations) to 0.8, 2.5 and 3.6 million, respectively. While PCIT has considerably reduced the number of gene-gene correlations based on objective criteria and therefore enriched a potential co-expression network for direct relationships, the resulting matrix sizes are still too large to serve as input for Cytoscape (see discussion). A further reduction of matrix sizes is therefore required, with a correlation strength threshold chosen to reduce the (largest) adjacency matrix to a size that can be accepted by Cytoscape when run on a 4-core MacPro with 9Gb of RAM). This reduction is here achieved by retaining a gene-gene correlation pair only if the absolute correlation coefficient $|r| \geq 0.9$. Unfortunately, this also removes lower-valued correlations that still may be significant as identified by PCIT (a perceived advantage of that approach). Table 6.1 summarised the initial number reductions.

Table 6.1 Adjacency matrix size reduction steps

	IFN-γ time course	**MCMV time course**	**IFN-γ + MCMV time course**
Initial data set	204	204	204
After 1st filtering	3.2	8	8.5
After PCIT	0.8	3.6	2.5
Cytoscape input	0.07	0.16	0.13

This table states (unit=millions) the number of pair-wise gene correlations that make up an adjacency matrix (a square matrix of Pearson correlation coefficients measured for every possible combination of two genes' linear similarity of expression profiles across all time points in a time course data set). For example, if the full input of 14299 genes are used for correlation, this results in 14299^2=204,461,401 correlation coefficients computed. First filtering refers to removal of genes that are not highly correlated (|r|>= 0.95) with at least one other gene (excluding itself). PCIT removes gene-gene pairs based on their statistical independence from a common third gene. "Cytoscape input" refers to a further necessary filtering step to reduce the adjacency matrix to a size that can be handled by Cytoscape, every time-course only precedes with those gene-gene correlations better than |r|>=0.9 (i.e. the threshold is the same for all three time courses and based on the necessary size reduction for the largest set).

With identical (even if arbitrary for correlation thresholds) criteria applied to each time course, early indications are the IFN-γ treatment is different from the other two conditions, affecting (reducing) the number of observed gene-gene similarities.

6.5.2 Gene cross-correlation maps show structural differences between the three time-courses

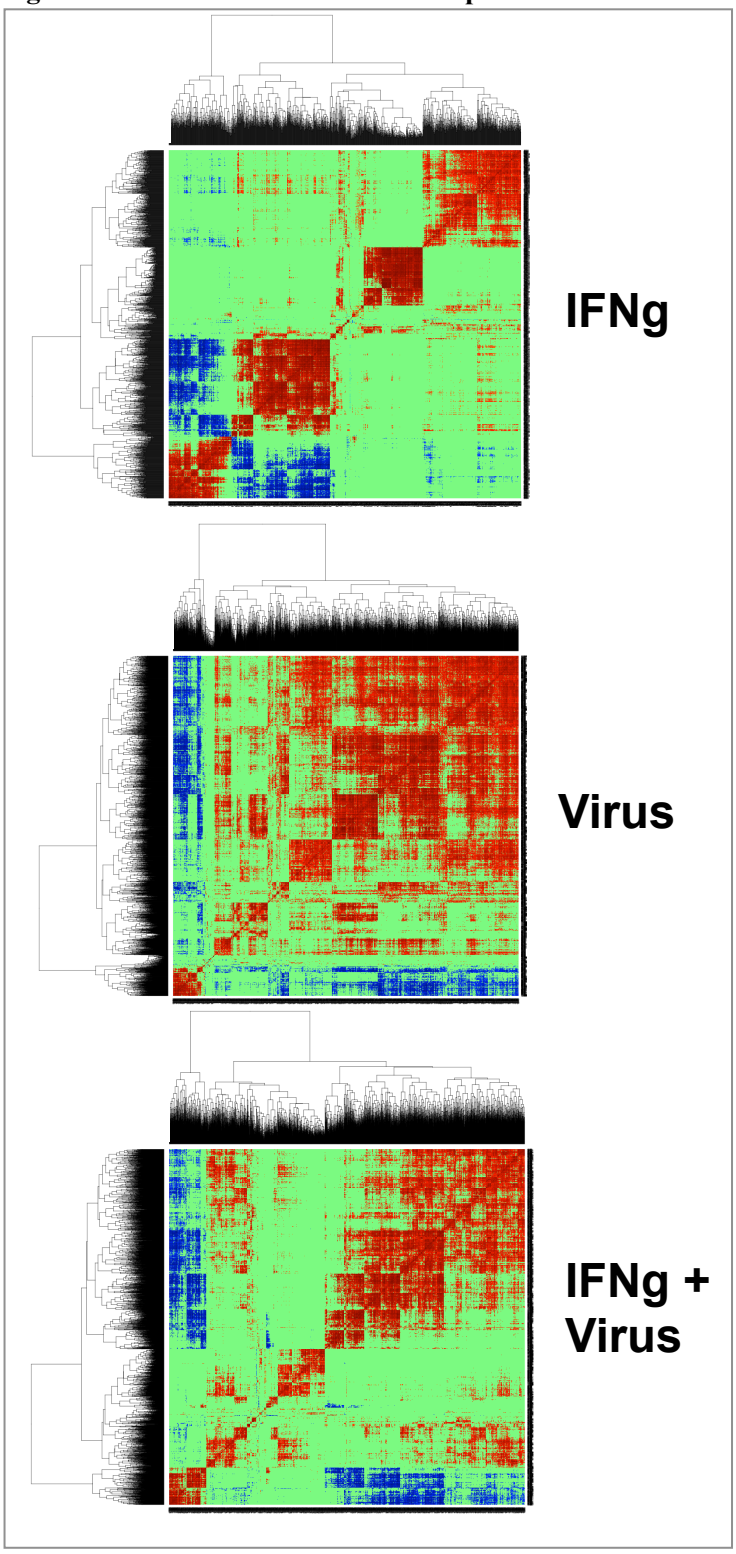
Prior to exporting data to Cytoscape (and thus reducing the size of the adjacency matrices), a useful visualisation of the existing gene correlation matrices (after 1st filtering and PCIT) is a simple plot of all correlation coefficients, as shown in figure 6.4 below. It should be noted that gene-gene correlations highlighted in green are those manually set to 0 because they did not pass the PCIT identification of significant correlations. As already clear from table 6.1, IFN-γ has notably fewer gene-gene correlations, but the visually observable differences also suggest that IFN-γ affects the coordination of genes that can be seen in virus-infected

macrophages. In IFN-γ treated macrophages, fewer genes are co-expressed, but in distinct gene sets. The combination of virus infection and IFN-γ treatment would appear to display an intermediate scenario. Although this cannot be assumed to be of statistical or biological significance, it gives cause to speculation (not tested in this thesis) that the virus is actually coordinating gene expression, and that this can be disrupted by IFN-γ. Taking all three outcomes into account, this implies there could be intrinsic differences between initial IFN-γ receptor signalling (IFN-γ treatment) and initial toll-like receptor signalling (on virus recognition). These may be due to induction or inhibition of different downstream genes (and pro- or anti-inflammatory cytokines) but technical shortcomings of viral infection cannot be discounted. That is, at Multiplicity-Of-Infection 1.0, work within the Ghazal group (data not shown) has shown that 30-40% of macrophages do not go on to express viral genes, which will affect measured mRNA transcription on the full macrophage population.

Also of interest here are negative gene-gene correlations, that is genes with a given expression trajectory that is seen inverted in other genes. In virus-infected macrophages, these can be described as a small number of genes negatively correlated with a large number of other genes. In IFN-γ treated macrophages, a small number of genes is negatively correlated with a small number of other genes.

In interpretation it is important to keep in mind that adjacency matrices (and later, networks) may portray visually compelling structures but lack numerical indicators for their significance.

Figure 6.4 Gene cross-correlation maps



This graph visualises the gene adjacency matrices for each time course. Rows and columns represent the same set of genes and are sorted by a clustering vector derived from genes' correlation coefficients. Each graph pixel is the Pearson correlation coefficient for a pair of genes. The diagonal indicates self-self correlations of a given gene. Pixels in the red colour spectrum are positively correlated, pixels in the blue spectrum are negatively correlated, green pixels are non-significant correlations as defined by PCIT, and set to 0

here. The maps contain 0.8, 3.6 and 2.5 million correlation coefficients for IFN-γ, Virus and IFN-γ+Virus, respectively.

A closer look at structural properties is obtained by transforming the gene correlation matrices into gene co-expression network graphs, representing gene similarity by edges drawn between gene nodes, and the overall layout informed by strength of correlation, number of correlated genes and other parameters useable in the visual arrangement of nodes. As a point of procedure, a perceived expression level increase for many genes at the 90 minute time point (which could in theory be biological fact or technical artefact) was separately tested for its influence on network structure, but the removal of this time point had no discernible effect on the resulting graphs (data not shown) or outcomes discussed here.

6.5.3 Cytoscape gene co-expression networks show structural differences between the three time courses as well as particular gene "hubs"

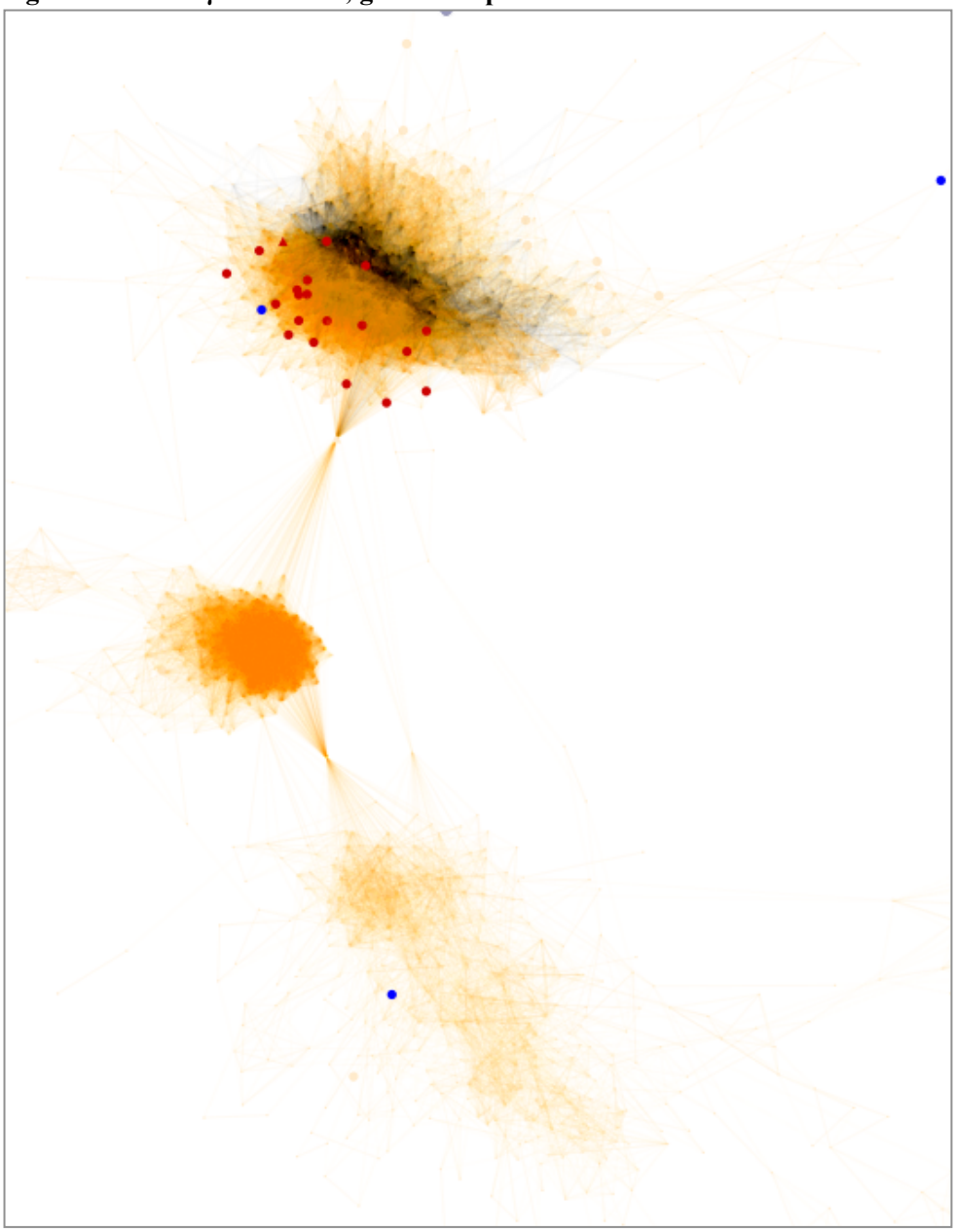
To examine gene co-expression patterns in greater detail, topographical gene co-expression networks generated in Cytoscape (figures 6.5a-c) are based on a layout-algorithm that determines the visual organisation of genes (nodes) through the supplied correlation coefficient information. The IFN-γ time course exhibits a remarkably different network structure, with three distinct "subnetworks" separated by a few "hub" genes that have similarities to genes in two subnetworks but do not fully match expression profiles in either. Interpreted at the highest level, these networks are (naturally) similar in information content to the structure revealed in figure 6.4, with virus-infected macrophages resulting in a dense network of closely correlated genes. If topological layouts are assumed to have a basis in the underlying biology (and this cannot here be confirmed), it could be suggested that this infection either coordinates gene expression of this set, or does not affect homeostatic coordinated gene transcription at all (although the latter is less likely given that the combined infection+treatment time course displays an in-between state that would not exist if virus infection had no effect on macrophages). In contrast, in interferon-gamma treated macrophages this tight

network is disrupted, with treatment causing genes to follow different expression trajectories over 12 hours. The actual expression profiles (data not shown) of genes in these subnetworks are not prominent in themselves (e.g. slow expression increases with some variability), which suggests that applying a standard clustering algorithm (Partitioning-Around-Medoids, PAM) would likely result in ignoring these gene sets in favour of those with strong expression changes, in effect favouring expression level over gene-gene relationships.

6.5.4 Cytoscape gene co-expression networks show alignment with meta-analysis results

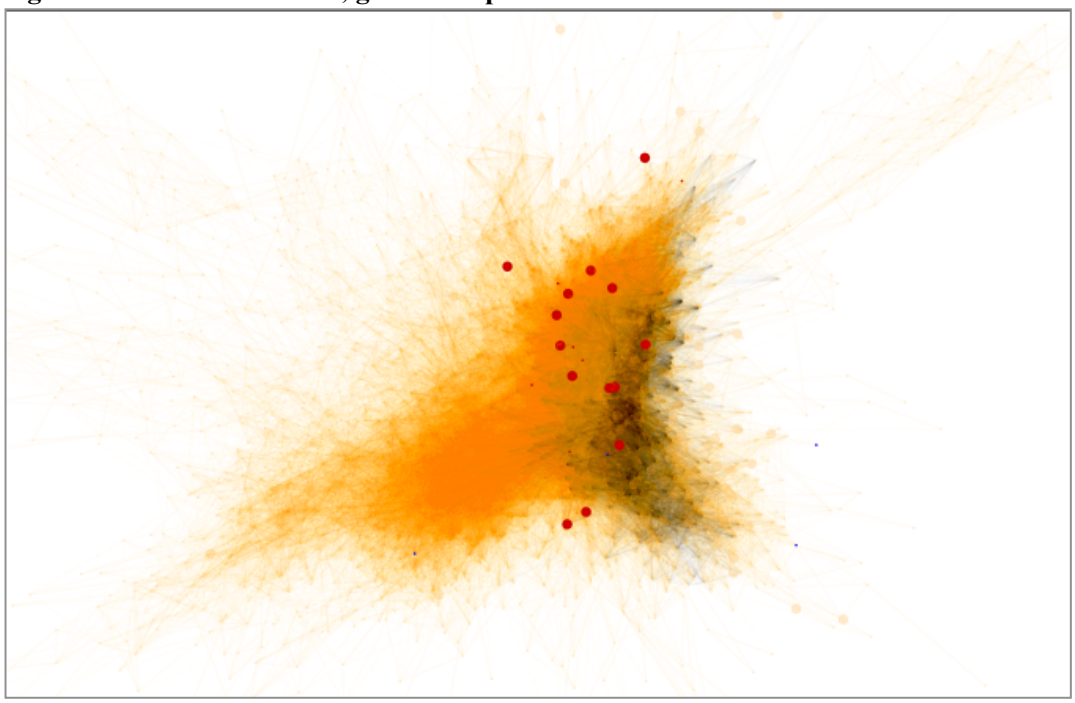
Highlighted gene nodes (red = up regulated by IFN-γ, blue = down regulated by IFN-γ) on figures 5a-c identify significant meta-analysis results that overlap with the MITCH12 time course microarray study. The limited number of highlighted genes is solely caused by genes not passing the correlation strength limits set by the network analysis. That is, where meta-analysis may identify an individual gene of consistent but low transcriptional changes, network analysis may exclude them because their expression is not sufficiently similar or independently correlated to another gene. All IFN-γ up regulated genes (n=20) are part of the immune response subnetwork, with down regulated genes (n=3) not stringently associated with any single subnetwork. Discussed in chapter 5 and amongst the notable matches between meta-analysis and network analysis are Stard3, Cd47, M6pr, Clic4, Lrp10, Tbk1, Mxd1 (all up regulated) and Ptpre (down regulated). For these genes, it would appear sensible to consider co-expressed genes into any formulated hypotheses. Notably, many of these genes (e.g. Stard3, Clic4, Tbk1, Lrp10) maintain strong co-expression (to other genes) when non-activated or activated macrophages are challenged with cytomegalovirus (figures 5b and 5c). In these other two biological conditions, expression trends are roughly similar to that of the IFN-γ activation alone, but they are by no means identical across the entire 12-hour period. In terms of interpretation, maintaining co-expression levels with other genes does not indicate that expression levels or patterns are not affected by the biological condition in question.

Figure 6.5a IFN-γ treatment, gene co-expression network



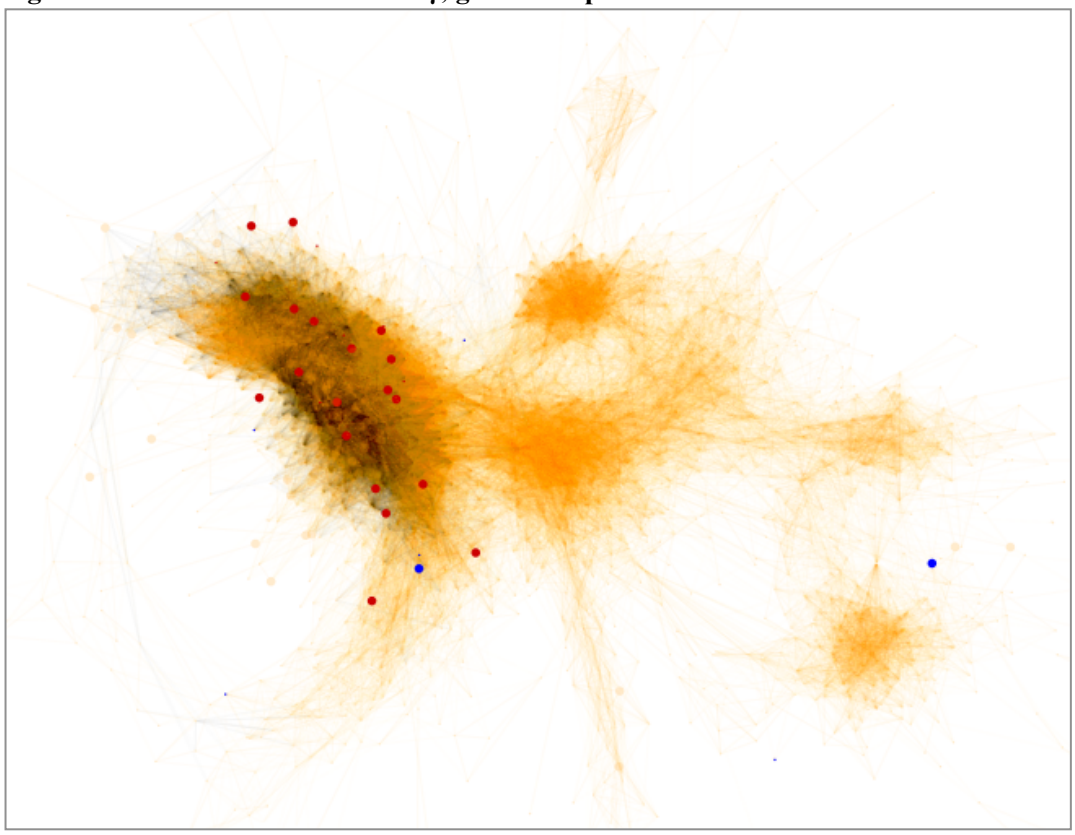
Gene co-expression network generated with Cytoscape (v2.8.0), using as input ~70,000 gene-gene correlation coefficients. Nodes are genes, orange edges between nodes represent high positive correlation between the two connected genes, black edges represent high negative correlation. Highlighted nodes are genes identified as significant in meta-analysis (red=up regulated by IFN-γ, blue = down regulated by IFN-γ).

Figure 6.5b Virus-infection, gene co-expression network



Gene co-expression network generated with Cytoscape (v2.8.0), using as input ~160,000 gene-gene correlation coefficients. Nodes are genes, orange edges between nodes represent high positive correlation between the two connected genes, black edges represent high negative correlation. Highlighted nodes are genes identified as significant in meta-analysis (red=up regulated by IFN-γ, blue = down regulated by IFN-γ).

Figure 6.5c Combined virus+IFN-γ, gene co-expression network



Gene co-expression network generated with Cytoscape (v2.8.0), using as input ~130,000 gene-gene correlation coefficients. Nodes are genes, orange edges between nodes represent high positive correlation between the two connected genes, black edges represent high negative correlation. Highlighted nodes are genes identified as significant in meta-analysis (red=up regulated by IFN-γ, blue = down regulated by IFN-γ).

6.5.5 Cytoscape gene co-expression networks may indicate different biological functionality.

Using IPA (Ingenuity Pathway Analyzer), the gene sets contained in the major IFN-γ subnetworks were tested for significant overrepresentation of gene ontology annotation categories. The uppermost subnetwork is significantly enriched for immune response, stress response and cell division genes, the middle subnetwork is enriched for general cellular processes, and the lowermost subnetwork for G-Protein coupled receptor activity, olfactory receptor activity and transmembrane receptor activity. Given the larger number of cross-correlated genes in the viral time course, enrichment is less specific and mainly limited to protein binding events. The time course with combined virus infection and IFN-γ treatment has

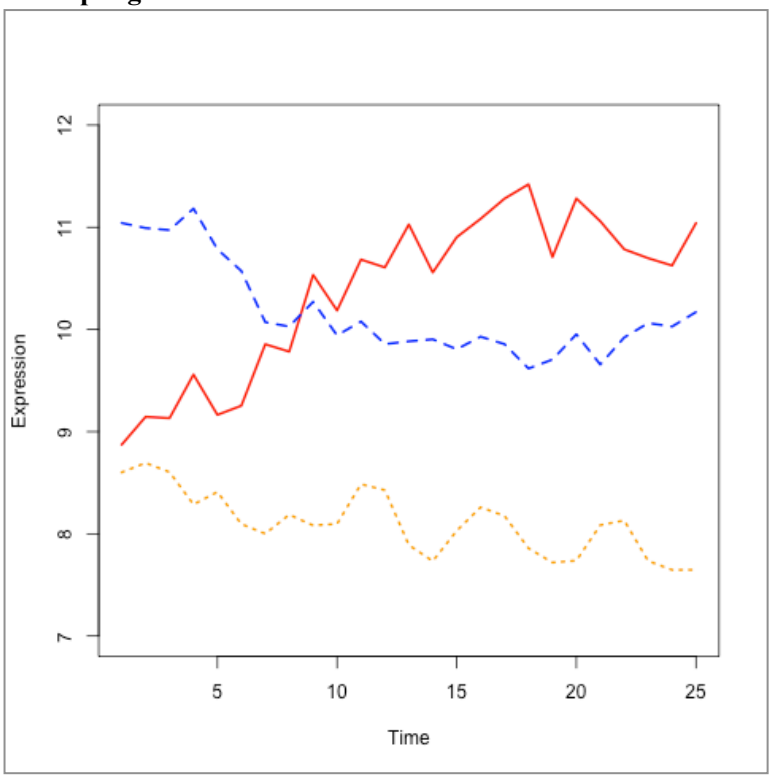
two distinct subnetworks that match IFN-γ, with the two small rightmost subnetworks enriched for G-Protein and olfactory receptor activity. These results are here listed only as general pointers of interest, since interpretations of large gene co-expression networks are not free of subjective bias.

6.5.6 Trafd1 (alias Fln29)

Trafd1 *is anti-correlated to sterol biosynthesis genes***.** Recently, sterol metabolism has been shown to be part of an intrinsic component of interferon signalling (Blanc 2013). In this context, the single node that in IFN-γ treated macrophages connects the uppermost subnetwork ("immune response") to the middle subnetwork ("cellular processes") is Trafd1 (of the multiple probes representing this gene, the most highly expressed is the one in question here). While it would be an over-interpretation to assume that Trafd1 has an actual biological coupling function between two distinct gene sets (because co-expression does not automatically imply biological function), their distinct location in the topographical layout motivates a closer examination.

Trafd1 It is positively correlated to 86 genes in the middle subnetwork and 138 genes in the upper subnetwork, but of particular interest are 19 genes with which it is negatively correlated in the upper subnetwork. Simple examination of these genes flags up *Sqle* and *Sc4mol*, both of which are participants in the sterol biosynthesis pathway and interact with SREBP2. This generates enhanced interest in *Trafd1* because research within the Ghazal group (Blanc 2013) had identified a likely link between an interferon induced antiviral response (mediated through STAT1) in macrophages and the synthesis of 25-hydroxycholesterol (which acts downstream of the above two genes and proteins). However, *Trafd1* was not known in this context. Its 12 hour expression profile in the IFN-γ time course is presented in figure 6.6 and includes the expression profiles of *Srebf2* (activator of cholesterol synthesis) and *Srebf1* (an isoform less specific to cholesterol).

Figure 6.6 Trafd1 and Srebf2 expression profiles on IFN-γ treatment of macrophages



Following IFN-γ treatment, 12-hour expression profiles of *Trafd1* (red), showing anti-correlation to *Srebf2* (orange, dotted) and *Srebf1* (blue, dashed). X-axis shows 25 half-hourly time points (starting at 0 hours and ending at 12 hours), Y-axis shows normalised absolute $\log_2$ scale expression level.

Expanding the interest in anti-correlation to the full sterol pathway (Table 6.2, 25 genes) shows that there is a general tendency for down regulation (indicated by anti-correlation with up regulated *Trafd1*) of these genes, even if they do not pass the $|r| \geq 0.9$ threshold set for the Cytoscape network.

Table 6.2 Trafd1 correlations to sterol pathway genes

Gene Symbol	**GeneName**	**Correlation to Trafd1 (IFN-γ time course)**
Acat1	acetyl-Coenzyme A acetyltransferase 1	-0.06
Hmgcl	3-hydroxy-3-methylglutaryl-Coenzyme A lyase	0.25
Hmgcs1	3-hydroxy-3-methylglutaryl-Coenzyme A synthase 1	**-0.86**
Hmgcr	3-hydroxy-3-methylglutaryl-Coenzyme A reductase	-0.54
Mvk	mevalonate kinase	-0.08
Pmvk	phosphomevalonate kinase	-0.01
Idi1	isopentenyl-diphosphate delta isomerase	-0.67
Mvd	mevalonate (diphospho) decarboxylase	-0.53
Fdps	farnesyl diphosphate synthetase	-0.54
Ggps1	geranylgeranyl diphosphate synthase 1	0.70
Ggps1	geranylgeranyl diphosphate synthase 1	0.72
Dhdds	dehydrodolichyl diphosphate synthase	-0.53
Fdft1	farnesyl diphosphate farnesyl transferase 1	**-0.84**
Sqle	squalene epoxidase	**-0.91**
Lss	lanosterol synthase	-0.63
Tm7sf2	transmembrane 7 superfamily member 2	-0.02
Ebp	phenylalkylamine Ca2+ antagonist (emopamil) binding protein	0.06
Sc5d	sterol-C5-desaturase (fungal ERG3, delta-5-desaturase) homolog (S. cerevisae)	-0.72
Dhcr7	7-dehydrocholesterol reductase	**-0.88**
Dhcr24	24-dehydrocholesterol reductase	-0.47
Dhcr24	24-dehydrocholesterol reductase	-0.78
Cyp51	cytochrome P450, family 51	0.13
Sc4mol	sterol-C4-methyl oxidase-like	**-0.90**
Nsdhl	NAD(P) dependent steroid dehydrogenase-like	-0.06
Hsd17b7	hydroxysteroid (17-beta) dehydrogenase 7	-0.66

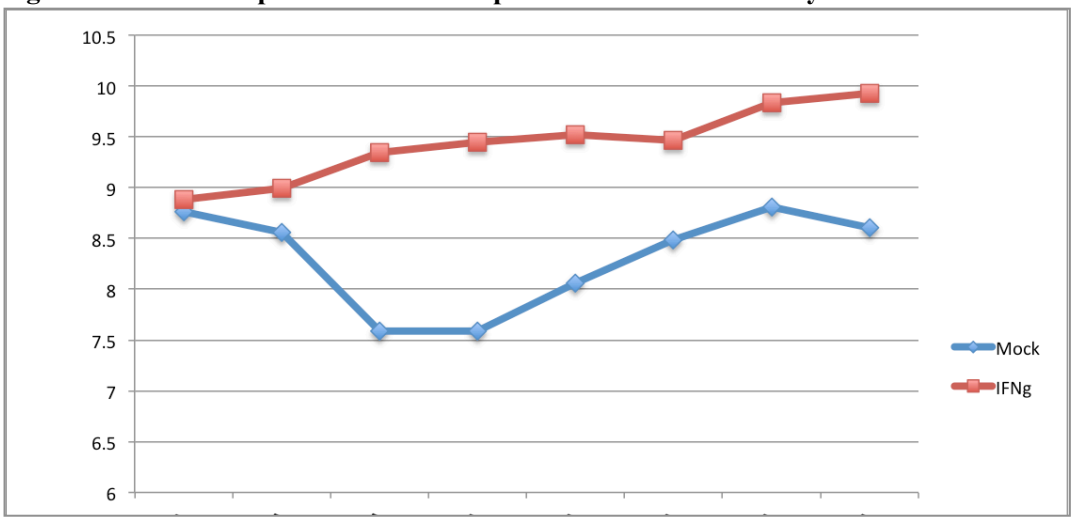
This assessment of negative correlation promotes further investigation of *Trafd1*. Firstly, its expression in IFN-γ stimulated macrophages is experimentally verified, passively through an existing other microarray time course study and actively through qPCR and siRNA experiments carried out in the Ghazal group for this purpose. Secondly, existing knowledge on *Trafd1* is summarised.

Trafd1 significant in multiple microarray studies. *Trafd1* is significantly (to several decimal places, and by any statistical test) up regulated by IFN-γ in all four microarray studies that form part of the meta-analysis in chapter 3. It is therefore also significant in meta-analysis, but not included amongst results in chapter 5

because its significance in individual studies means it does not match the criterion of meta-analysis-only discovery.

Trafd1 up regulation confirmed in independent time course study. A study of the effects of IFN-γ on macrophages was carried out within the Ghazal research group (subsequently to and independently of MITCH12), utilising a different microarray platform and experimental protocols. The retrieved expression profile for Trafd1 (figure 6.7) suggests 2-fold up regulation (or higher, if considered relative to a mock condition) over a period of 7 hours, broadly confirming the trajectory observed in MITCH12.

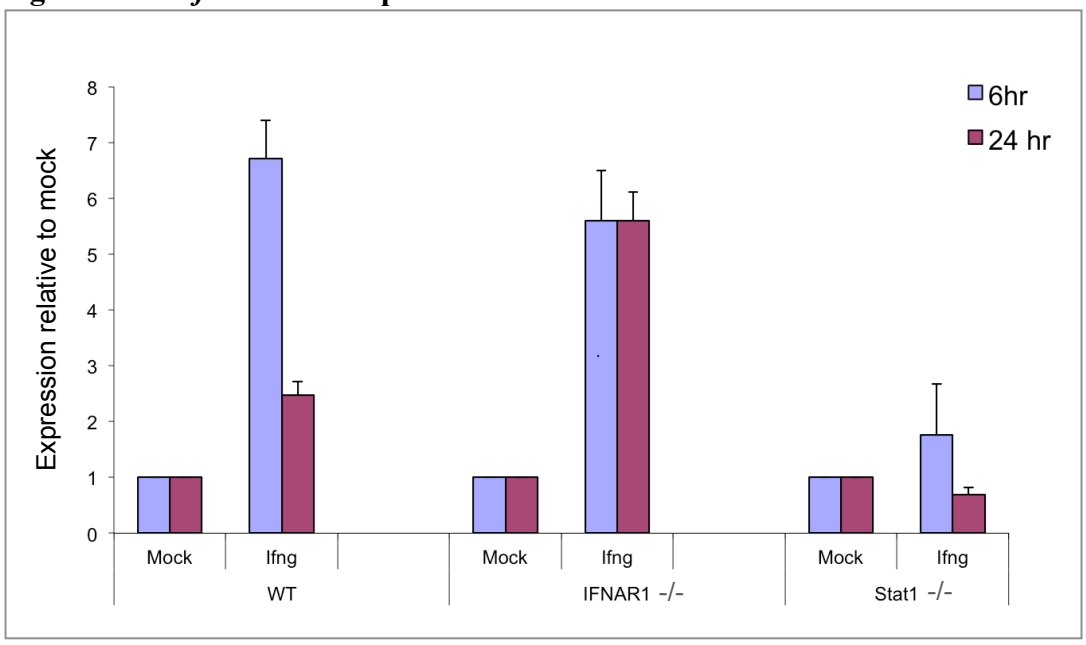
Figure 6.7 Trafd1 expression in an independent time course study



Graph of the absolute log2 scale expression levels (Y-axis) of *Trafd1* in a time course of IFN-γ treated murine bone marrow derived macrophages. Y-axis represents time periods measured, ranging from 0-30 minutes at the first X-ordinate, to 420-450 minutes at the last X-ordinate.

Trafd1 up regulation confirmed through real time quantitative PCR. A quantitative real time PCR experiment carried out (in triplicate) in validation of *Trafd1* confirms up regulation by IFN-γ in wild type mouse macrophages, where this is further established by IFN-γ not being up regulated in Stat1 knock-outs. Interferon alpha receptor (IFNAR) knock-out does not limit IFN-γ up regulation of *Trafd1*.

Figure 6.8 *Trafd1* mRNA expression validation

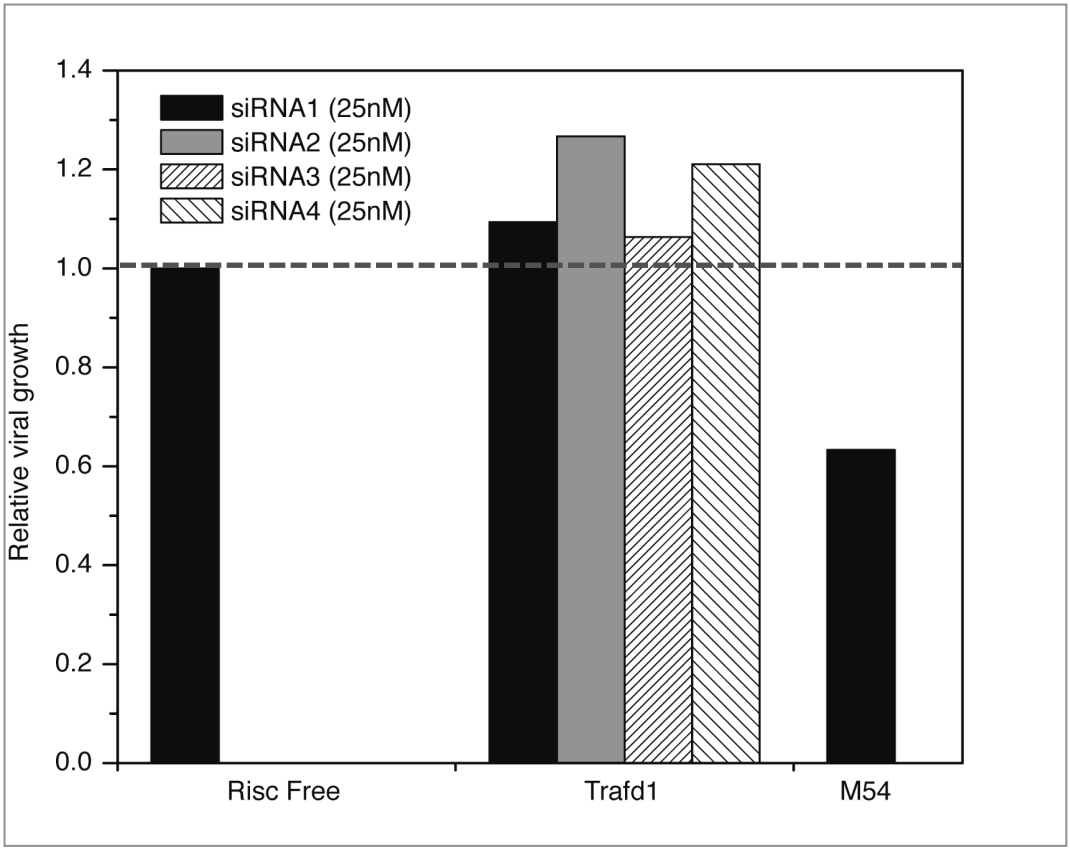


Graphs show levels of mRNA expression of Trafd1 in response to IFN-γ treatment (100u/ml), shown in relation to mock treatment. Each treatment was measured at 6 and 24 hours in bone marrow derived macrophages (BMDM) cultured from either wild-type mice, interferon alpha receptor (IFNAR1) knock-out mice or Stat1 knock-out mice. Bars represent the means ± standard error of the mean of n=3 biological replicates (=mice). These experiments were performed using Real Time quantitative Reverse Transcription PCR (qRT-PCR) based on TaqMan Applied Biosystems, with experiments run and analysed by Mathieu Blanc within the Division of Pathway Medicine. The TaqMan qRT-PCR and the treatment protocol used has been described previously (Blanc, Hsieh et al. 2011) and only differs in the use of the gene tested, which is here Trafd1 (TaqMan primer assay ID: Mm00551535_m1).

siRNA assay of Trafd1 suggests but does not confirm anti-viral effects. A validation study of *Trafd1* knocked down (figure 6.9) by four different siRNAs (each done in triplicate) shows increased viral growth relative to RISC control, suggesting that *Trafd1* may have a direct or indirect role in restricting viral growth, although the strength of this evidence is not beyond doubt. This may be due to this type of siRNA assay lacking sensitivity to detect increases rather than decreases, with observations (within the Ghazal group) up relative increase normally below 1.5 for any experiment, and the validation is considered

incomplete until a larger number of replicate runs (10-12) are performed beyond the three shown here.

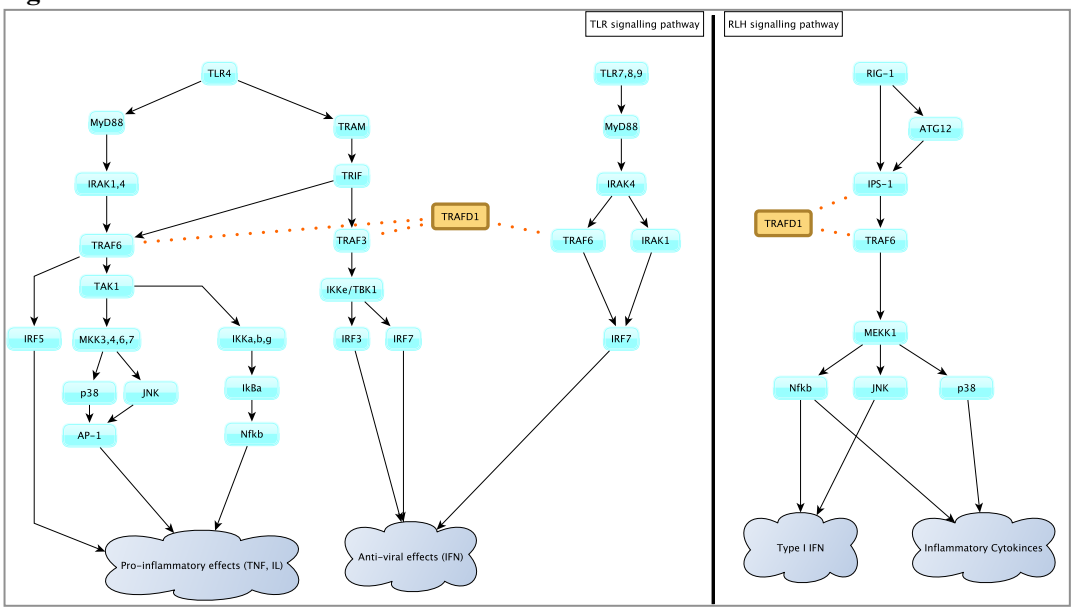
Figure 6.9 *Trafd1* laboratory validation in siRNA knock-down



Experiment data provided by Wayne Hsieh, Ghazal group. Four different (in terms of sequence) siRNA probes are used to knock down Trafd1 in order to observe the effect on viral growth. Each knock down is performed in triplicate, but error bars are not added to this graph because it has not reached the final experiment size of 10-12 replicates. Construction of the GFP-encoding Murine Cytomegalovirus (MCMV-GFP, originally named: pSM3fr-rev) used in this study was previously described (Angulo, Ghazal et al. 2000, Angulo, Ghazal et al. 2000). The virus was propagated in mouse NIH-3T3 fibroblasts. For Trafd1 experiments, 1.5×10^4 primary murine embryo fibroblasts cells (pMEF) in 80 µl of DMEM 10% CS medium lacking antibiotics was added to the siRNA:Dharmafect 1 complexes. Growth medium was removed and 100 µl of the siRNA: Dharmafect 1 liposomes were added. 48h post transfection, the transfection media was first removed. Cells were then infected with GFP tagged murine cytomegalovirus (MCMV-GFP) [multiplicity of infection (MOI = 0.05] for 1h at 37° C. Virus replication was monitored as a function of eGFP fluorescence hours post-infection (24-72h for MCMV) using the POLARstar OPTIMA plate reader (BMG Labtech). Virus replication slopes over the linear phase were calculated and normalised to Risc Free transfected cells, and the mean replication was calculated. M54 siRNA was used as a positive control as it targets MCMV polymerase. Data represent the mean of 3 replicates per siRNA.

Known or proposed Trafd1 functions. *Trafd1* is suggested to be a negative regulator of the innate immune response and modulation of inflammatory function in macrophages. Specifically, Trafd1 is known as interferon-inducible and LPS-inducible, with its induction STAT1-dependent (Mashima, Saeki et al. 2005). Sanada et al (Sanada, Takaesu et al. 2008) identified Trafd1 as a negative feedback regulator of TLR, suppressing TLR4-mediated NF-kappaB activation by binding to TRAF6. Others (Richards and Macdonald 2011) also suggest binding to TRAF3, but in either case the binding interrupts anti-viral signalling, with the actual mechanism unknown. A summary of this information is provided in figure 6.10. There appears to be no known function of Trafd1 (or its binding partner Traf6) relating to sterol biosynthesis.

Figure 6.10 Schematic overview of known Trafd1 function



Pathways drawn with yED graph editor. Dotted lines indicate binding with the effect of inhibition of the targeted signalling pathway, although the mechanisms remain unknown.

Other "hubs". Based on the amount of supporting evidence available through annotation, biological interest in *Trafd1* has here outweighed interest into that of the other subnetwork-connecting genes in the IFN-γ network, and these are *Olfr1116* and *Wdfy2*. For these, little is known about their functionality, although the protein of Wdfy2 has been shown to interact with Foxo1 protein (*Foxo1* is a transcription factor) for a putative function in promoting adipogenesis (Fritzius

and Moelling 2008) and *Foxo1* has been identified as an inhibitor of cellular antiviral responses and IFN-β (Lei, Zhang et al. 2013). If their status as "hubs" is of similar relevance to that of *Trafd1*, validation of this functionality would require further research.

6.6 Discussion

Network analyses. As already recognised in the early days of microarray technology (Quackenbush 2003) and discussed to this day (Gillis and Pavlidis 2012), guilt-by-association networks have some promise but are far from optimal in their ability to reliably identify biologically significant (rather than statistically or numerically significant) gene networks. Much of this is related to experiment design, which often does not capture true biological variation because the number of samples or biological conditions is too small. The selection of dis(similarity) metric is also of importance, and Pearson correlation was here only chosen on the basis of it having been the focus of a supercomputing-based collaboration. Time-course studies have an advantage in that their interpretation is more straightforward (coordinated gene expression over time) than that for gene-association analyses in studies with multiple biological conditions (if a gene maintains correlation in different biological settings, what does this signify). However, major drawbacks remain in form of assigning biologically significant function to genes or gene sets, requiring thorough follow-up work to verify results.

Gene filtering. The addition of PCIT to the workflow is an attempt to reduce the final gene co-expression networks to only those genes that have a direct expression association and excluding those associations that can be readily explained by a common third gene. This approach does not ensure a causal-links-only gene network, but should be enriched for it. In theory, this is an improvement over simply choosing gene-similarity thresholds (in this case, correlation coefficients) with the assumption that all indirect gene-gene associations will have lower correlation levels. PCIT allows "weaker" gene-gene associations to remain in the data as long as they are statistically proven to be

independent of a third gene, making for a less objective threshold. In practice, the effect cannot be easily assessed on unknown data (i.e., outside simulations on data sets with known outcomes), and the improvements achieved may be diminished by a need for further massive reductions in data size due to computational limits. In this case, PCIT decreases gene-gene correlation matrices by a factor of around 2 to 3 (table 6.1), but these considerable reductions still exceeds Cytoscape's computational limitations on the available hardware. The therefore necessary arbitrary reduction to gene-gene correlations $|r|\geq 0.9$ excludes a large number of low but significant gene-gene associations.

New hypotheses related to Trafd1. Prior research (Blanc, Hsieh et al. 2011, Blanc 2013) has identified that the TLR signalling, interferon type I and type II signalling pathway may interact with the sterol metabolic, expanding the role of lipogenic transcription factors (SREBP-1, SREBP-2) from nutrient level regulation to responses to bacterial and viral infection. The evidence in this chapter is insufficient to suggest that *Trafd1* has a causal relationship to reduction of sterol synthesis that is parallel to its role as negative regulator of TLR4 signalling to prevent runaway innate immune responses. It is also unclear if the reverse may be a possibility, with reduced sterol synthesis leading to up regulation of *Trafd1* and therefore reduction of TLR4-mediated responses. The (verified) evidence obtained in this chapter is limited to the statements that *Trafd1* is induced by interferon-gamma mediated JAK-STAT signalling and that its induction and mRNA increases strongly (and coincides in terms of time points) with decreases in *Sqle* and *Sc4mol* (also Srebf1 and Srebf2) gene transcription. Taken together with Ghazal group findings on STAT1 being the link between interferon signalling and sterol biosynthesis, this warrants prospective experiments to confirm a role for Trafd1 beyond negative regulation of LPS-signalling.

Comprehensiveness of analysis. This chapter provides only an isolated analysis, further analyses have been carried out (expression thresholding, fuzzy clustering, statistical tests for gene periodicity) and are planned (statistical test for time-dependent gene expression changes). These have identified separate aspects (genes

involved in autophagy, genes potentially related to circadian rhythms, genes related to vesicular transport) that do not overlap with findings discussed in this chapter. The network analysis itself is also not fully comprehensive, in this chapter favouring the development of a particular targeted hypothesis over a complete listing of results, which includes complete sets of enrichment tests performed on subnetworks (using IPA or DAVID). This chapter also omits network-specific metrics regarding degree of node-connectedness, average path lengths, other methods of gene hub identification etc., mainly because interpretation of such outcomes cannot easily be associated with biology. The use of Cytoscape for generating network graphs is an additional limitation in the analysis workflow used here. Other R-based[42] or external[43] network graph tools may allow for higher maximum node or edge numbers, enabling relaxation of the here used final step of gene-gene correlation strength ($|r| \geq 0.9$).

Value of network analyses. While this chapter shows that explorative methods (cluster and network analyses) can be successful in identifying broad biology (here, global effects of IFN-γ treatment) and potentially interesting candidate genes (here, *Trafd1*), they remain very subjective in their use. That is, a different set of parameters and values (different gene-gene distance metrics, correlation thresholds, processing algorithms) may lead to equally valid but entirely different new findings and hypotheses. In this instance, care was taken to choose processing steps on the basis of computational limitations rather than on the basis of biological expectations, avoiding user bias. The work in this chapter was initially (before result interpretation) independent of the Ghazal groups work on sterol synthesis and results were therefore not guided towards these outcomes. Result interpretation is inevitably biased towards prior domain knowledge, which is unavoidable for "omics" type data, where algorithm-based (explorative or statistical) prioritisation of genes still produces sizeable gene sets that require either the prior domain knowledge mentioned above or sufficient means to investigate each constituent gene. Network analysis has here also demonstrated

[42] http://www.bioconductor.org/packages/release/BiocViews.html#GraphsAndNetworks
[43] www.graphviz.org

that it may in some cases surpass the usefulness of clustering approaches like K-means or PAM, which do not highlight the role of individual genes and would not have called for further investigation of the mildly interesting expression profile or the set of genes anti-correlated to it. However, this does not mean network analyses can be taken at face value. It is in fact risky to do so, given that a gene co-expression network's topographical layout is dependent on small differences that may not represent true biology because there are not sufficient numbers of samples or replicates to reliably recognise small differences. Additionally, any gene-gene correlation coefficients can be due to both (causal) biology or (non-causal) random matches.

Conclusions. Viral infection and IFN-γ activated macrophages appear to be distinct in their genome-wide coordinated responses, with the level of viral infection used here possibly controlling or negating expression of a large proportion of macrophage genes (resulting in high co-expression numbers). IFN-γ activated macrophages present an image of modular coordination, with immune response genes forming a separate subnetwork from genes with other or more generic functions. When infection is measured in activated macrophages, a visually intermediate scenario presents itself, and if representative of actual biology, this may come closest to a naturally occurring response (activated macrophages dealing with infection). While these are not unexpected findings as such, the results here encourage further work on detailed content of subnetworks identified in the IFN-γ and combined treatment/infection regimes. However, this work is outside the scope of this chapter and would also benefit from analysis methods with a less global approach, e.g. statistical analyses of the differential regulation between the treatment regimes, using specific gene sets of interest only. *Trafd1* as a single gene hypothesis is attractive for further research, as it fits into (but does not stem from) a current research theme within the Ghazal group, and also shows some promise in identifying a further factor in the crosstalk between IFN-γ (which Trafd1 is not known for but here obviously present) and TLR signalling (which Trafd1 is known for as a negative regulator, but with direct TLR signalling or LPS treatment not used in this study).

With the above limitations in mind, an objectively applied (without expected outcomes in mind) network analysis can identify promising and otherwise unnoticeable biological hypotheses to pursue. Unlike statistical analyses they do not provide probabilistic results or predetermined significance thresholds. Combined with the above-mentioned risk of too much confidence in the quality or meaning of gene-gene correlations, successful outcomes require strong prospective experimental testing of new hypotheses derived from those results.

Chapter 7

Discussion

This research set out to investigate if the use of statistical meta-analysis techniques on a limited collection of small microarray studies can identify otherwise masked transcriptional changes in the IFN-γ activation of macrophages. Although the newer technology of high-throughput next generation sequencing is set to replace microarray technology in time, microarray meta-analysis remains an important subject given the continuing use of microarray technology and the availability of large numbers of legacy microarray data sets in public repositories. Additionally, despite a wealth of research into macrophage signalling pathways and interferon-stimulated genes by microarray and other techniques, the field remains open for statistically powered solutions that are able to detect weaker but important genes involved in macrophage type II interferon signalling or by extension for other biological hypotheses such as a viral challenge in type II interferon signalling.

The investigation of the primary research hypothesis above was converted into a framework of four research questions aimed to provide an increased understanding of which genes are involved in the type II interferon response in macrophages and how to detect them.

- Firstly, can statistical meta-analysis provide meaningful biological results when applied to a heterogeneous (in terms of experiment design, not subject matter) and restricted set of small microarray studies?
- Secondly, can existing meta-analysis models be improved?
- Thirdly, can the potential scope of microarray meta-analyses be increased by imputing gene expression values that are missing due to merging of different microarray platforms?
- Finally, can the existing body of knowledge with regard to IFN-γ-mediated JAK-STAT signalling in macrophages be expanded based on meta-analysis results?

7.1 Research conclusions

Taken together, this research has demonstrated that statistical meta-analyses are a suitable and improvable tool for primary identification of gene candidates in type II interferon signalling in macrophages, and that with the aid of biological domain knowledge these can be developed into novel biological hypotheses.
Relevant findings are outlined below on the basis of the above stated research questions. Details on these are provided as part of section 7.2.

Can statistical meta-analysis provide meaningful biological results when applied to a heterogeneous and restricted set of small microarray studies?
All evidence obtained in chapters 3, 5 and 6 supports a positive answer to this research question. Chapter 3 provides quantitative evidence that a statistical meta-analysis results in gene transcription discoveries not made in any individual studies and also points to clear advantages for non-parametric methodology (Rank Product meta-analyses and Fisher's meta-analysis based on Rank Product tests). Chapters 3 and 5 corroborate that this quantitative improvement in terms of numbers of significant results can be biologically verified, both broadly against lists of interferon-stimulated genes in any experimental setting, and specifically against independent macrophage experiment data. A detailed investigation of final meta-analysis result lists provides clear and novel candidate hypotheses to be tested in a macrophage activation system.

Can existing meta-analysis models be improved?
Improvement of models was shown to be of utility for a set of heterogeneous microarray studies. Quantitative evidence suggests that the alternative weighting scheme (proportional study weighting) for the Rank Product meta-analysis shows favourable performance with increasing stringency of the alpha criterion (figure 3.6), that is, lowering of the p-value significance threshold enables this method to identify more true positive genes than other methods. This trend is less clear for

the equal study weighting alternative, but both suggested methodology advances are an improvement over the original Rank Product meta-analysis, where this interpretation must currently be limited to cases in which the selected microarray studies are very unbalanced in terms of sample size. Moreover, in the selection of per-study analysis methods when applying Fisher's combination of probabilities, to a heterogeneous data set, a Rank Product test should be favoured over a parametric test.

Can imputation of missing gene expression values increase the scope of microarray meta-analyses?

Remarkably, while not very sensitive or useful to obtain estimates of biological expression, statistical estimation of large numbers of gene expression measurements (which are missing not due to random errors or caused by other variables, but due to not all genes being represented on all microarray platforms) is more successful at recovering known statistical outcomes than analysis with missing values or, obviously, not analysing these genes at all.

Can the existing body of knowledge with regard to IFN-γ-mediated JAK-STAT signalling in macrophages be expanded?

Meta-analysis has identified (depending on model) 70 to 107 IFN-γ up regulated and 66 to 175 IFN-γ down regulated genes with expression too low or variable to be identified in any individual microarray study. Subsequent aggregation reduced this list to 106 up regulated and 152 down regulated genes, and corroboration against independent data sources specified priority lists for experimental validation, consisting of 58 up regulated and 47 down regulated genes, previously not known or strongly confirmed to have a role in type II interferon signalling in murine macrophages. In confirming better than random overlap between meta-analysis results and known biology, biological and bioinformatical follow-up has highlighted several gene transcription hypotheses open to prospective testing. Many of these concern possible crosstalk between type I and type II interferon signalling (*Tbk1*, *Ikbke, Clic4*), between type II interferon signalling and IL-10 signalling (*Batf, Ptpre*), are of interest in linking type II interferon signalling to

sterol biosynthesis in the macrophage (*Stard3*, *Pgrmc1*, *Itga4*, *Obfc2a*, *Lrp10*, *Ehd1*), or provide further possible insights in the regulation of between immune responses by the circadian clock (*Csnk1e*).

7.2 Implications of research outcomes

This research has implications at the level of statistical methodology as well as macrophage biology, and these are here discussed in context of existing knowledge.

In terms of statistical methodology, this research is supportive of prior research (Rhodes, Barrette et al. 2002, Choi, Yu et al. 2003, Hong and Breitling 2008) in that the combining of multiple studies provides notionally more power to identify differential expression differences in case-control microarray studies. Where the cited studies have assessed this outcome mostly quantitatively and with only a very limited amount of biological validation, the research here has provided both numerical and biological evidence for the relevance of meta-analysis.

Research findings here also support the conclusion (Hong and Breitling 2008) that non-parametric meta-analyses (Rank Product or Fisher's model based on rank statistics) have advantages over parametric methodology (effect size models, Fisher's model based on parametric test results) when using a small heterogeneous set of microarray studies. Earlier cases made for the use of parametric models (Choi, Yu et al. 2003) only find agreement here in terms of absolute numbers (in that any meta-analysis improves on single-study analyses), but not in terms of relative numbers when compared to non-parametric approaches. However, it is unknown how the Bayesian model proposed by Choi et al would fare in comparison to non-parametric models given the data sets used here.

The research here also suggests that one particular improvement to Rank Product meta-analysis is beneficial over the standard model in cases of large study size heterogeneity. The here proposed proportionally weighted Rank Product meta-

analysis has not been suggested before, and partially this may be due to previous research (Rhodes, Barrette et al. 2002, Choi, Yu et al. 2003, Hong and Breitling 2008) having pragmatically or intentionally limited their real, standard or simulated microarray data sets to approximately balanced study sizes only; in these scenarios, similar study sizes mean the issue is less relevant. Clearly, as evidenced by the microarray studies available for type II interferon activation of macrophages, the issue exists, and the alternative weighting scheme proposed should be a primary choice in such cases.

Large-scale imputation of a particular type of missing gene expression values in microarray meta-analyses has not been subject to research before, presumably because the notion of generating gene expression estimates for genes that are not included on a given microarray platform is not intuitive. However, the large number of genes which have to be omitted from a meta-analysis (or are subject to reduced power) for this reason have driven an actual investigation into this issue and found that in terms of statistically significant results, imputation recovered a larger proportion than simply accepting missing values (figures 4.5 and 4.6), with BPCA or a simple sample-wise regression leading to the best imputation estimates. This usefulness does not extend to reliability of differential expression measurements using imputed values. While showing more promise than expected, it requires dedicated further research to establish if this approach works well for microarray data sets with other characteristics than the ones used here, and for that reason its application to the meta-analyses performed in this thesis was considered premature.

Meta-analysis has identified genes that are consistently regulated by IFN-γ in macrophages, with their characterisation for example pointing to a role in **sterol metabolism.**

Concurrent but independent work within the Ghazal group of STAT1-mediated links between interferon response and sterol biosynthesis (Blanc, Hsieh et al. 2011, Blanc 2013) suggests that the association of these genes with IFN-γ signalling should motivate prospective testing of the following hypotheses.

Up regulation of Stard3 by IFN-γ suppresses lipogenesis. A role in cholesterol transport (Strauss, Liu et al. 2002), sterol biosynthesis (Kishida, Kostetskii et al. 2004) and repression of lipogenesis (Borthwick, Allen et al. 2010) has been previously identified outside an immune response context, but no literature currently supports its regulation by IFN-γ.

Down regulation of Pgrmc1 by IFN-γ inhibits sterol biosynthesis via SCAP and Insig-1. As a direct interaction partner of SCAP and Insig-1 (Suchanek, Radzikowska et al. 2005), this gene is strongly associated with sterol biosynthesis and promotes sterol biosynthesis in some cells (Hughes, Powell et al. 2007), but this has not been shown in macrophages and no literature currently supports its regulation by IFN-γ.

Up regulation of Lrp10 by IFN-γ is mediated by SREBP2 protein. Evidence exists for *Lrp10* to be indirectly mediated by SREBP2, although this has only been shown for human hepatic cells (Huang, Zhou et al. 2008) and it assumes that Lrp10 is a subfamily of LDLR protein as has been shown for humans (Brodeur, Theriault et al. 2012). No literature currently supports the (independently validated) meta-analysis finding of a relation to IFN-γ signalling.

Up regulation of Ehd1 by IFN-γ regulates cholesterol uptake. Independent research suggests that *Ehd1* regulates cholesterol uptake in fibroblasts and is involved in the transport of vesicles containing Rab11 (Naslavsky, Rahajeng et al. 2006, Naslavsky, Rahajeng et al. 2007), where Rab11 itself is known to promote phagocytosis in macrophages (Cox, Lee et al. 2000) and meta-analysis here suggests its up regulation by IFN-γ.

Meta-analysis has also identified genes that are consistently regulated by IFN-γ in macrophages, but where literature evidence only points towards type I interferon signalling, suggesting possible **crosstalk between type I and type II interferon** signalling.

Up regulation of Tbk1 and Ikbke by IFN-γ causes a negative feedback loop enhancing type I interferon signalling over type II interferon signalling. *Tbk1* is a known inducer of IFN-β and IRF3 through TLR signalling and type I interferon signalling (Hemmi, Takeuchi et al. 2004, Perry, Chow et al. 2004). The meta-

analysis finding of *Tbk1* regulation by IFN-γ is corroborated through a matching up regulation of *Ikbke*, where this gene has previously been found to be functionally equivalent to *Tbk1* (Ikeda, Hecker et al. 2007). *Ikbke* has also been found to be a crosstalk regulator between type I and type II interferon signalling by phosphorylating STAT1 and preventing the formation of STAT1 homodimers and therefore shifting gene transcription towards type I interferon signalling (Ng, Friedman et al. 2011), while no such hypothesis has as yet been proposed for *Tbk1*.

Clic4 is up regulated by IFN-γ and involved in macrophage deactivation through IRF3. *Clic4* has been proposed as a regulator of macrophage deactivation in response to nitric oxide induction and LPS or LPS and IFN-γ signalling (He, Ma et al. 2011, Malik, Jividen et al. 2012). Through phosphorylation of IRF3 it may positively regulate LPS signalling, although the results of meta-analysis suggest LPS activation is not required as a primary trigger.

Meta-analysis has further highlighted genes related to **IL-10 signalling**.

Down regulation of Ptpre by IFN-γ induces or prevents inhibition of JAK-STAT signalling through phosphorylation. An isoform of the protein has been found to inhibit (in myeloid cells) IL-6 or IL-10 induced JAK-STAT signalling by preventing phosphorylation of STAT3. While meta-analysis suggests down regulation by IFN-γ, the above paper finds *Ptpre* not involved with STAT phosphorylation through type I or type II interferon responses.

Batf up regulation by IFN-γ directly or indirectly inhibits IL-10 production. In human macrophages, AP-1 (transcription factor required for IL-10 production) can be inhibited by BATF protein (Deppmann, Thornton et al. 2003) and it has been shown that IFN-γ also suppresses AP-1 (Hu, Paik et al. 2006).

Meta-analysis has identified a potential role for IFN-γ in regulating **molecular clocks**.

Csnk1e is down regulated by IFN-γ and directly or indirectly regulates the molecular clock in macrophages. The circadian or molecular clock has been found relevant to macrophages in their regulation of immune responses (Hayashi,

Shimba et al. 2007), with anti-viral responses under circadian control via JAK-STAT signalling in human T-cells (Bollinger, Leutz et al. 2011). *Csnk1e* is a circadian clock gene associated with PER2, and although TNFa and IL-1b have been shown to suppress clock genes including Per2 in mouse fibroblasts (Cavadini, Petrzilka et al. 2007), no previous evidence supports the meta-analysis result of *Csnk1e* down regulation by IFN-γ.

Network analysis of type II interferon signalling has shown meaningful overlap with meta-analysis results and emphasized biological importance and implications of **negatively correlated** gene expression patterns, here identifying a gene negatively correlated to **sterol biosynthesis** genes.

In an independent validation of meta-analysis results, an objective network-analysis of a microarray time-course study has shown that such data-driven methods (as opposed to hypothesis-driven meta-analyses) can lead to new biological insights despite a lack of probabilistic metrics and with explicit avoidance of bias in the various processing steps (i.e. not guiding results based on prior knowledge). In this case new biological insight has been identified in a topological network link (*Trafd1*) between two subnetworks as a possible direct or indirect connection between type II interferon signalling and the sterol biosynthesis pathway. The highlighting of *Trafd1*'s potential relevance to sterol biosynthesis is rooted in negative correlation patterns. Its general pattern of up regulation by IFN-γ over time is the inverse of that observed for two important sterol genes (*Sqle* and *Sc4mol*), and this negative correlation was obvious in a network graph but is unlikely to be found by other methodologies, although a complete listing of all highly anti-correlated gene pairs would of course contain it. In general terms, this should encourage closer investigation of strong negative correlation patterns within time course data, because related gene function may not always be indicated by positively correlated co-expression. In follow-up, a high proportion of other sterol pathway genes were also found to be negatively correlated (anti-correlated) to *Trafd1* (table 6.2), including the activator of cholesterol biosynthesis (*Srebf2,* figure 6.7). In conjunction with recent findings (Blanc 2013) of a likely link between an interferon induced antiviral response (via

STAT1) and synthesis of 25-hydroxycholesterol acting downstream of *Sqle* and *Sc4mol* genes and proteins, this resulted in a new hypothesis that *Trafd1* may be a relevant actor in this link. This hypothesis is refined by existing research on *Trafd1* as a negative regulator of TLR-mediated NF-kappaB activation (Sanada, Takaesu et al. 2008, Richards and Macdonald 2011) and inducible by interferon and LPS (Mashima, Saeki et al. 2005).

Although there are limitations, if this research is generalised to other biological hypotheses and slightly different data compositions, it should provide a strong incentive to utilise large amounts of free and publicly available microarray data for formal, automatic and cost-effective meta-analyses, where an additional level of biological follow-up work by the analyst or a close collaborator is strongly indicated in order to propose laboratory-testable hypotheses. Given the observed advantages in combining statistical results with biological validation efforts, the results obtained in this thesis should also encourage inter-disciplinary learning.

7.2 Limitations of research outcomes

The biological hypothesis and meta-analysis approaches used in this thesis are subject to narrow parameters, with any conclusions not necessarily transferrable beyond the specified treatment (IFN-γ), the mouse bone marrow derived macrophage system, meta-analysis model and data set characteristics. In the strictest sense, these limit the interpretation of this thesis to the following statement: "Rank-Product and proportionally weighted Rank-Product meta-analysis is clearly able to provide novel hypotheses for type II interferon signalling in murine macrophages if applied to a set of four differently-sized and Affymetrix-based microarray studies investigating the effect of a range of doses of IFN-γ measured at different time points in the activation of macrophages".

Study selection. Although any of the meta-analyses used in this thesis are easy to apply on a prepared set of microarray studies, particular challenges are posed in the correct identification of suitable studies, the matching of gene probes between

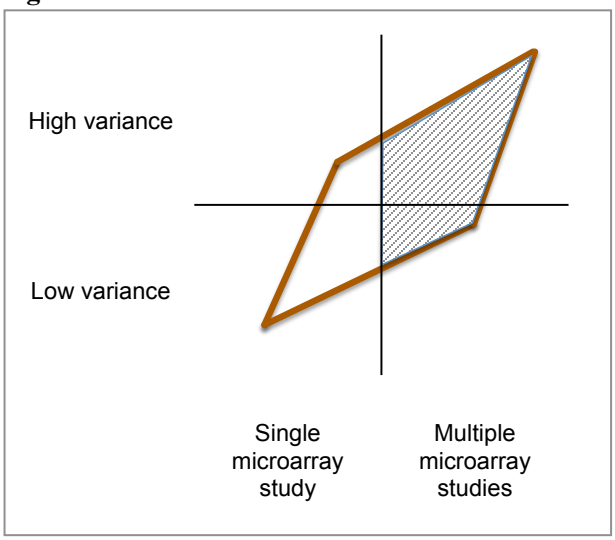
studies and the processing of individual studies. These concerns are particularly applicable if microarray data sets are based on different platforms, raising issues of different measurement scales (dual-colour hybridisations versus single-colour hybridisations), probe representations (single 30-70mer oligonucleotide sequences or multiple small 11-mer or small oligonucleotides), and single-study analysis strategies. Although chosen for their experiment design similarities, the level of such similarity is difficult to define and does make no attempt to test the implications of mouse background (C57Black/6 or Balb/c), macrophage activation or development state. While the broad question "what effect does IFN-γ have on murine macrophages" is answered, different conclusions would likely be obtained if study selection were either more or less specific.

Data simulation. Often, meta-analysis models (or any other new statistical approaches) are tested on "standard" data sets or simulated data. This was considered here, but those schemes were ultimately replaced by systematic biological assessments of results because a) standard data sets are only standards for historical reasons rather than purposefully generated, b) simulation would have to cover a prohibitively large parameter space (study size, variance, missing values, scale, proportions of high-fold-change-genes etc.) in order to truly generalise any results, and c) biological assessments can be seen as a reasonable compensation for the (albeit artificial) generalisation obtained through simulated or unrelated biological data. Here, the non-existent (for simulation) or very limited (for using standard data sets) biological validation is replaced by in-depth biological assessments of results in a global context (chapter 3) and in a system-specific (chapter 5) context. As a consequence, findings in this thesis cannot support the use of Rank-Product meta-analysis as a suitable choice for all small collections of microarray studies, but as a choice that has been shown to work well in this example.

Type II interferon signalling. While meta-analysis can provide a deeper understanding of IFN-γ initiated JAK-STAT signalling in macrophages, it is important to note that a complete characterisation of this pathway is not a goal in

the research carried out here. To address the main hypothesis of identifying small transcriptional changes otherwise masked by data variance in individual microarray studies, only those results were considered that are unique to meta-analysis, that is, genes that are already significant in any individual studies were excluded from interpretation (see figure 7.1). This does for example dismiss *Trafd1* (identified through network analysis) from discussion in chapter 5, as it is found to be significant in all of the individual microarray studies. However, if complete pathway characterisation is a goal and individual study results are included in the assessment of IFN-γ induced or regulated genes, limitations would still apply in that the studies used do not model the biological effects of other signalling pathways activated through pathogen recognition or through stimulation by other immune cells.

Figure 7.1 Genes of relevance to this thesis



The area enclosed in red represents (not to scale) genes identified through statistical testing (single-study inference tests or meta-analysis), where this number is dependent on data variance and number of studies. Microarray data with high gene expression variance increases difficulty to statistically identify differentially expressed genes, and a single microarray study has limited power to identify small expression changes with statistical significance. The shaded area contains those genes that are discovered only through the application of meta-analysis. The unshaded areas are statistically significant results obtained in single microarray studies and also form part of the response to IFN-γ macrophage activation, but their lower variance facilitates easier statistical identification in single studies and are therefore not suited to address the hypothesis underlying this thesis.

Meta-analysis significance. Although probabilistic analyses are more objective than machine learning approaches, ultimate results are still dependent on choice of models and parameters, which is here relevant in two ways. Firstly, it is a given that acceptance of a less stringent (e.g. p<= 0.05 instead of p<=0.01) false positive error threshold would have led to a larger number of genes identified as significant and may have produced further biological hypotheses. Secondly, the use of a multiple testing correction algorithm could also have performed the role of a more stringent alpha level, and would be recommended for citing significance levels in publication, but for comparing meta-analysis results this would have introduced a further element of complexity in that the correction methods commonly used differ for the three meta-analysis models, and would also differ for statistical testing performed in individual studies. Although not used here because the issue multiplies the outputs to be assessed, a separate investigation could be included in future research.

Sensitivity analysis. The definition of "combinable" studies to be used for meta-analysis is elastic and sensitivity analyses are a useful tool in standard meta-analysis theory to address this. Sensitivity analyses ask the question of "how robust is a meta-analysis with respect to different meta-analysis models, study size, study quality, or inclusion criteria". They are usually applied by repeatedly performing meta-analyses for different combinations or selections of the above parameters and then interpreting the spread of outcomes so obtained. Extending this concept to microarray meta-analysis in the particular instance of this thesis, meta-analyses could be repeated by permuting the number and size of studies considered, by including further studies that use other treatments (e.g. IFN-β), other cell types (e.g. fibroblasts, monocytes), other organisms (e.g. human, rat), or other platforms (e.g. Illumina arrays, custom-printed arrays). While this is desirable and the full set of studies already collected could make this technically feasible, it is extremely complex: A) sensitivity analyses have to be performed for each gene, B) in the absence of detailed biological interpretation (such as that done in chapter 5) it is not clear how to interpret a finding of more or fewer genes achieving a given (or changed) overall effect size, C) at the level of RNA

expression and with respect to the characteristics of different cell types or signalling pathways within a cell, the experiment factors listed above will likely defy simple explanations. With these issues perhaps making a sensitivity analysis for those biological factors impractical, a sensitivity analysis excluding different studies in turn could be feasible in order to detail the effect of study sizes and study numbers. The practical ability to do so technically exists after code solutions for meta-analyses and outcome assessments are implemented, but the additional effort in assessing this type of meta-analysis robustness was considered secondary to assessing the biological implications of results (chapter 5), where robustness is characterised by comparisons against known biology and independent experiments. Additionally, as a comparison of meta-analysis model outcomes is the purpose of chapter 3, section 3.4 is in effect a sensitivity analysis for the effect of using different meta-analysis models, although not named as such.

Network analysis. In contrast to statistical probabilistic models for testing inference, network analysis is a comparatively subjective tool, that is, depending on user choices for parameters and parameter values (primarily scale of data values, distance/similarity metric thresholds, type of distance/similarity metric, clustering algorithm) the obtained results may be meaningful but only represent one possible outcome out of many. Although this research attempts to use objective and consistent parameter values to avoid results fully guided by user-expectations, the obtained gene co-expression networks will represent only an incomplete picture of the biological effects measured in the system.

7.3 Future research

A major consideration of this thesis was the interfacing between different expertises, and this is planned to remain a future research focus in light of biological data becoming more rather than less complex. Future research therefore pursues both the promotion of a deeper understanding of the biological system through combining data from multiple sources, as well as improvements to meta-analysis methodology to facilitate this.

To this end, follow-up research to this thesis is planned or has already been initiated, as outlined below and going beyond intended publications of the proportionally weighted Rank Product model, biological results, network analysis, and (requiring further simulation work) missing data imputation. Future research on the basis of this thesis centers on follow-on experimental validation of results, the use of meta-analysis for biological hypotheses, enhancements and alternatives to meta-analysis models, and outreach to make basic meta-analysis approaches widely accessible.

Follow-up testing of biological hypotheses. Analyses in this thesis have identified several hypotheses (discussed in section 7.2) on transcriptional changes induced by type II interferon activation of macrophages. Firstly, collaboration with experimentalists will be sought for independent validation of induction by IFN-γ (partially already in progress in form of siRNA screens and real-time quantitative PCR). Secondly, funding will be sought to prospectively perform studies testing the newly identified hypotheses and assigning mechanisms of action, with particular preference for the linking of type II interferon responses in macrophages to sterol biosynthesis and type I interferon response. Thirdly, a later stage may include validation of and research into those genes that are significant by meta-analysis (tables 5.8 and 5.9) but are associated with little or no current functional annotation.

Meta-analysis and cloud computing. Research into the technological aspects of future application of meta-analyses to new or legacy microarray data sets has been initialised by the Edinburgh Parallel Computing Centre (Alastair Hume) and me. Initially, the research question will be addressed by an MSc student, with research question and guidance provided by me on the basis of this thesis: Can meta-analyses be remotely applied to microarray data sets stored in the "cloud" (remote, internet-accessible data and software storage)? If successful, further collaboration will be sought with microarray data set providers, with the aim of exploiting large amounts of new or legacy microarray data for simplified remote meta-analyses by non-statisticians.

Applied meta-analyses. Meta-analysis has shown great utility for gaining deeper insights into type II interferon signalling in macrophages, which permits expansion of this methodology to an expanded view of this system. Further research is therefore set to exploit the initial set of candidate studies collected for potential meta-analysis, comprising a total of 30 studies. These meta-analyses will be aimed to provide a different view of immune responses to that pursued in this thesis, and consist of a) responses to type I interferon, b) responses to full macrophage activation (both IFN-γ and LPS signal), c) responses in human macrophages, and d) responses by other immune cells than macrophages.

Network analyses. Although not a central hypothesis of this thesis, the presence of anti-correlated gene networks in the substantiation of meta-analysis results by network analysis has uncovered relevant biology and implies higher value than is currently acknowledged. In addition to negative correlation, secondary results (not included in this thesis) also pointed to limitations of (dis)similarity based networks by excluding part correlations (in the sense of breaking a time expression profile down into smaller time windows) that nonetheless share a biological relation. It is therefore planned to extend formal investigations (using alternative distance metrics and identifying probabilistic statistics for network relevance) into the identification and meaning of anti-correlated and part-correlated gene networks, using existing time course microarray data in the department.

Meta-analysis advances. This research has shown that improvements to existing meta-analysis models can be beneficial, and future research is set to build on this. Firstly, the proportionally weighted Rank Product meta-analysis will undergo further comparative testing and biological application within the plans outlined in the previous paragraph. Secondly, following refinements and further testing of missing value imputation, this approach will also be incorporated and introduced to the research community if successful. Thirdly, time courses are a frequent design with biological microarray studies, and a future focus will be their use for meta-analytical approaches, where this will require comparison to other data-integration methods. Lastly, high-throughput next-generation-sequencing technology is set to become a new standard for post-genomic experiments, and it will be a priority to investigate if and how meta-analysis can also be relevant in this context.

7.4 Final thoughts

Prior to this research, meta-analysis had been shown to have promise in combining microarray study results, but the lack of biological verification in statistical research, the lack of meta-analysis methodology comparisons in biological research, and the as yet incomplete understanding of transcriptional outcomes in type II interferon signalling have driven the research question and research methodology in this thesis.

The number, confirmed biological relevance and inclusion in prospective research of findings only attributable to microarray meta-analysis points towards it being a very useful research tool for even small sets of heterogeneous microarray studies. An obstacle to their use may be the identification of suitable microarray studies (which is difficult for analysts) and the merging of data and application of meta-analyses (which are difficult for biologists). Another obstacle may simply be that there is insufficient awareness in the biological research community of statistical meta-analysis models and their merits. This is only addressable by outreach activities (which this thesis and planned future work is part of), and can only have

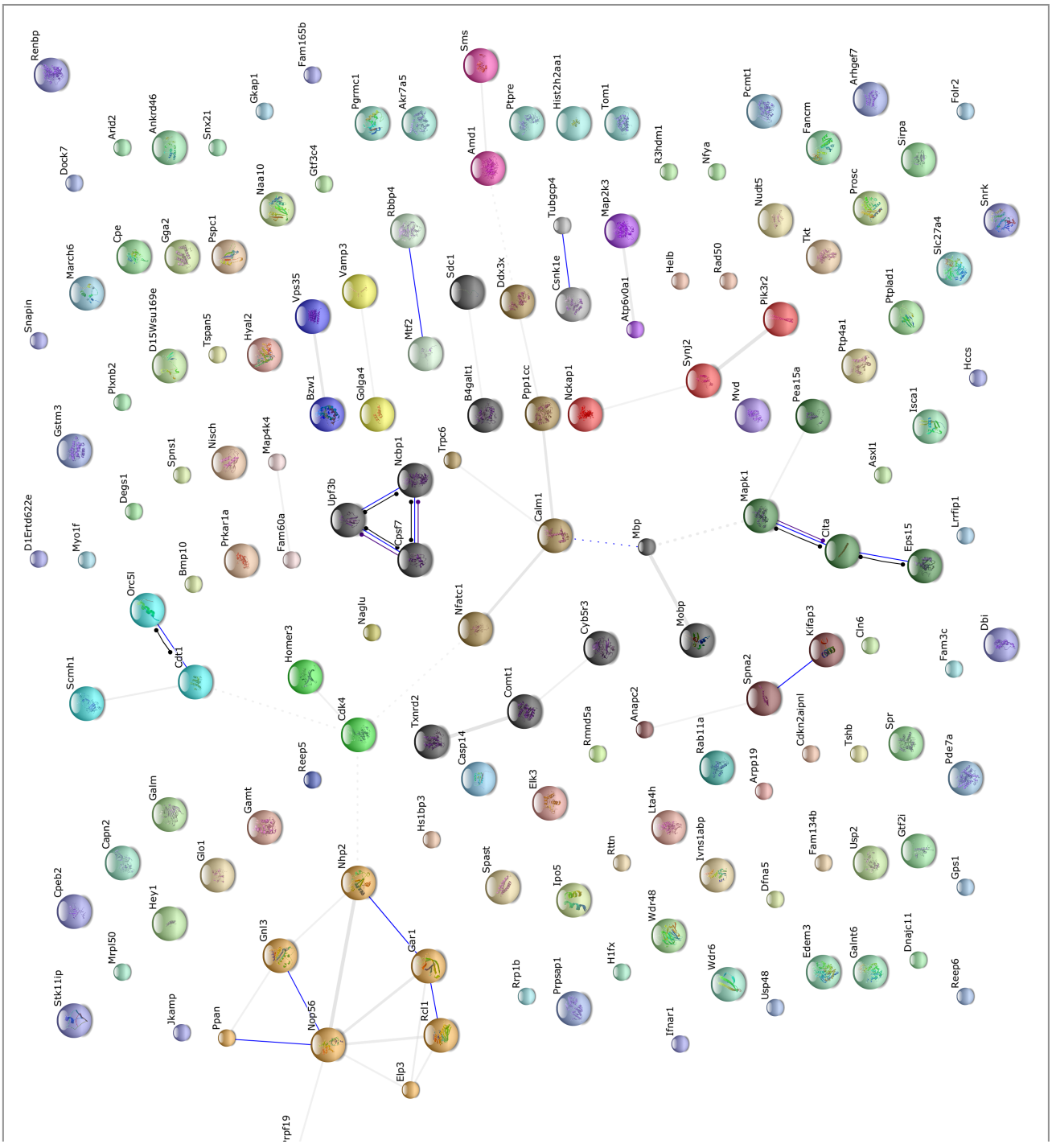
positive consequences for research, given the availability of very large numbers of publicly available microarray data sets that often complement one another.

There is also little doubt that the value of supplementing statistical methodology with biological domain knowledge is real, and that this can not only be achieved through close research collaboration, but perhaps maximised through professionals in one domain acquiring modest amounts of knowledge in the second domain.

Appendix

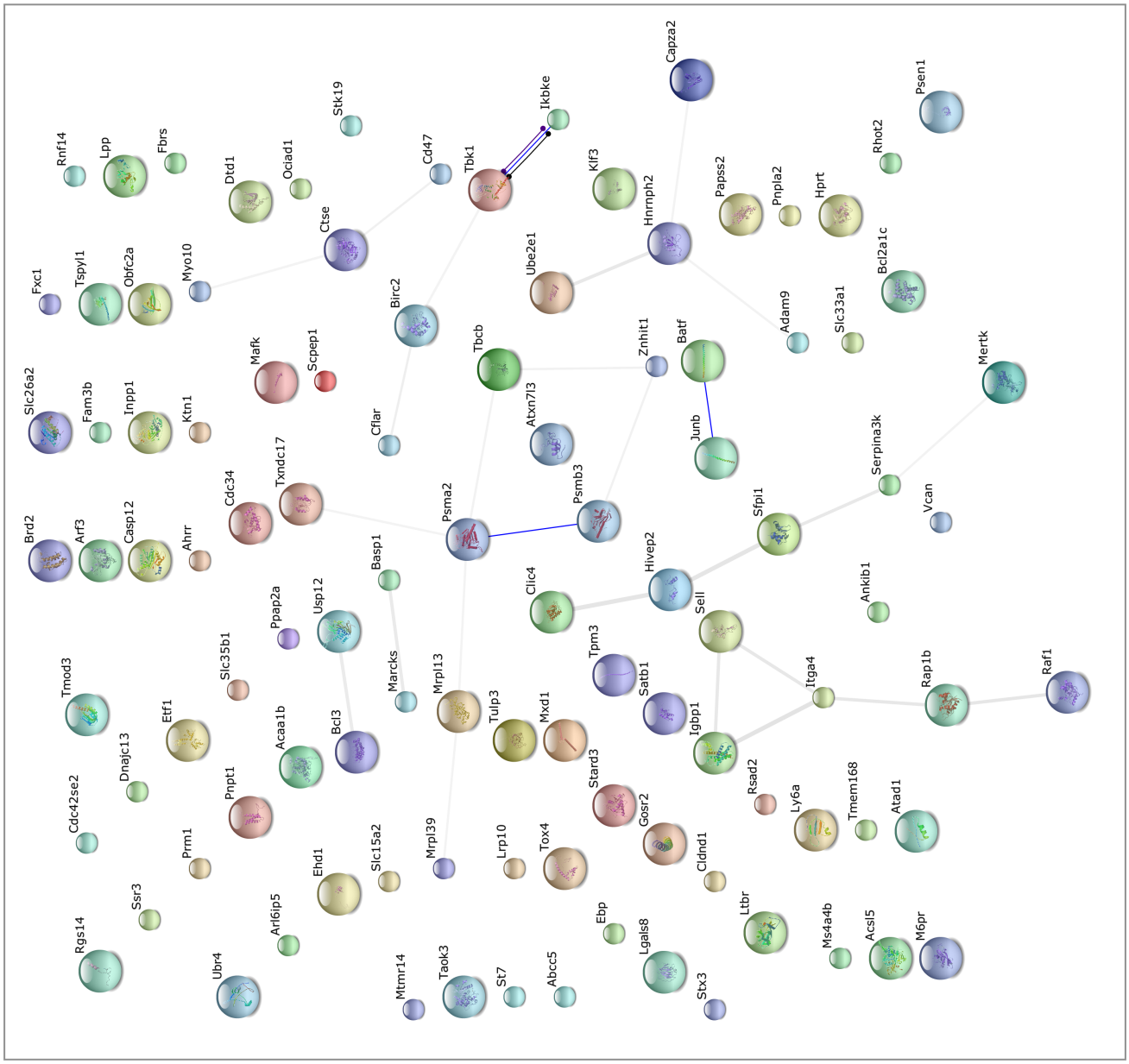
A1- Protein network (STRING), down regulated genes

(Based on genes that 2 out of 3 meta-analysis model families identified as significantly **down regulated by IFN-γ)**



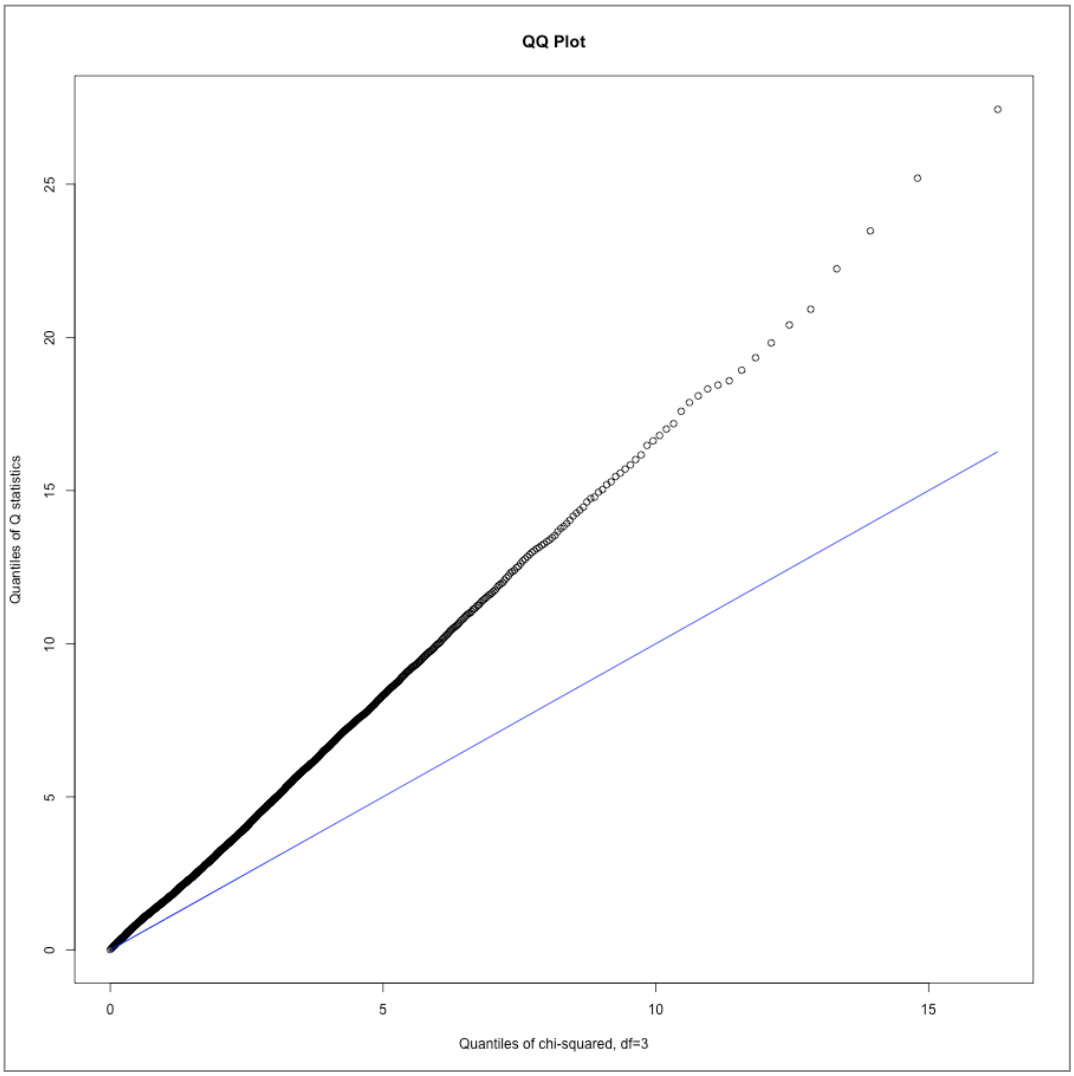
A2 - Protein network (STRING), up regulated genes

(Based on genes that 2 out of 3 meta-analysis model families identified as significantly **down regulated by IFN-γ)**



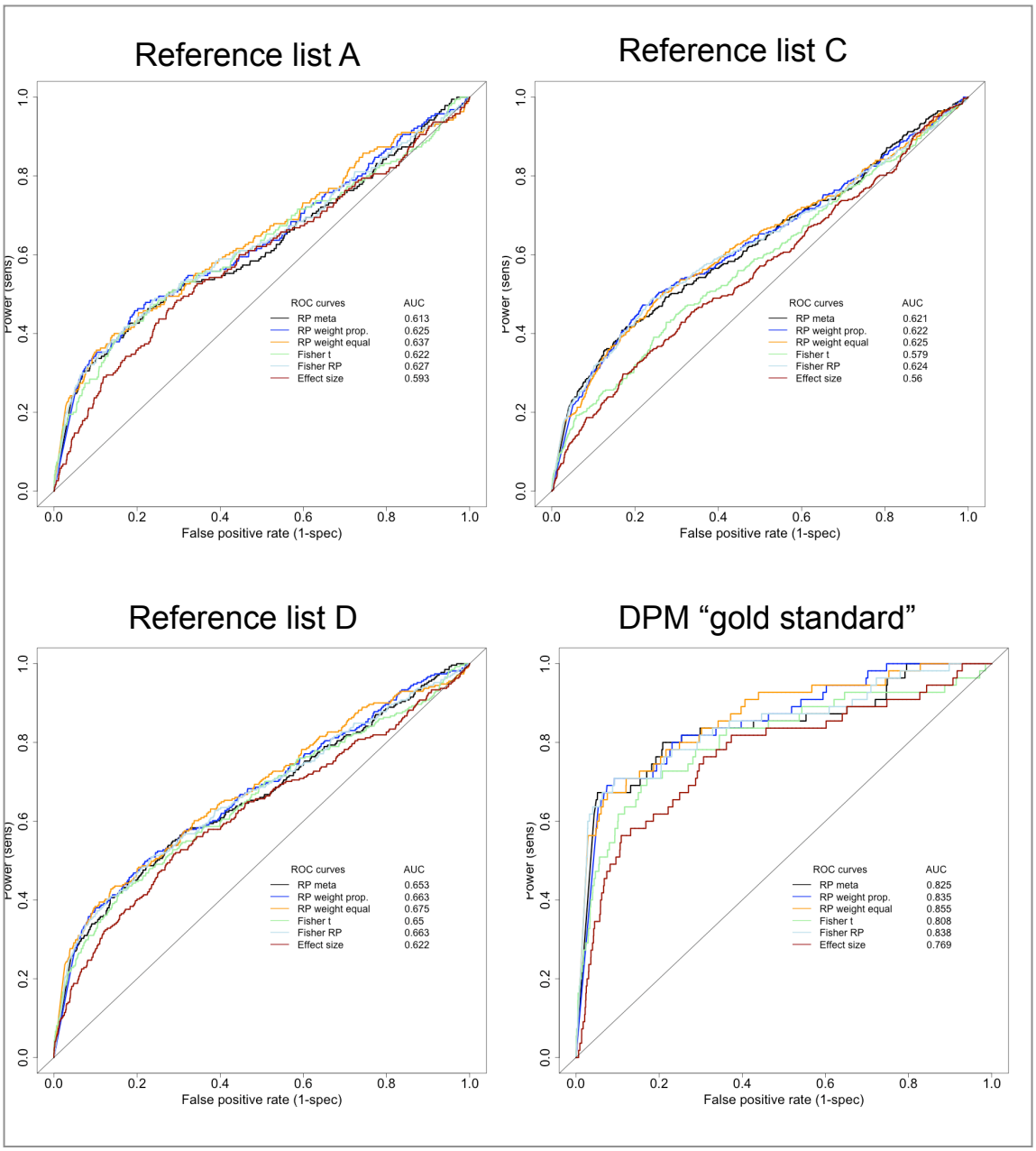
A3 – QQ plot for deciding on effect size model

Quantile-Quantile (QQ) plot of theoretical chi-squared distribution quantiles (degrees of freedom=3) against observed Q heterogeneity-statistic for each gene.



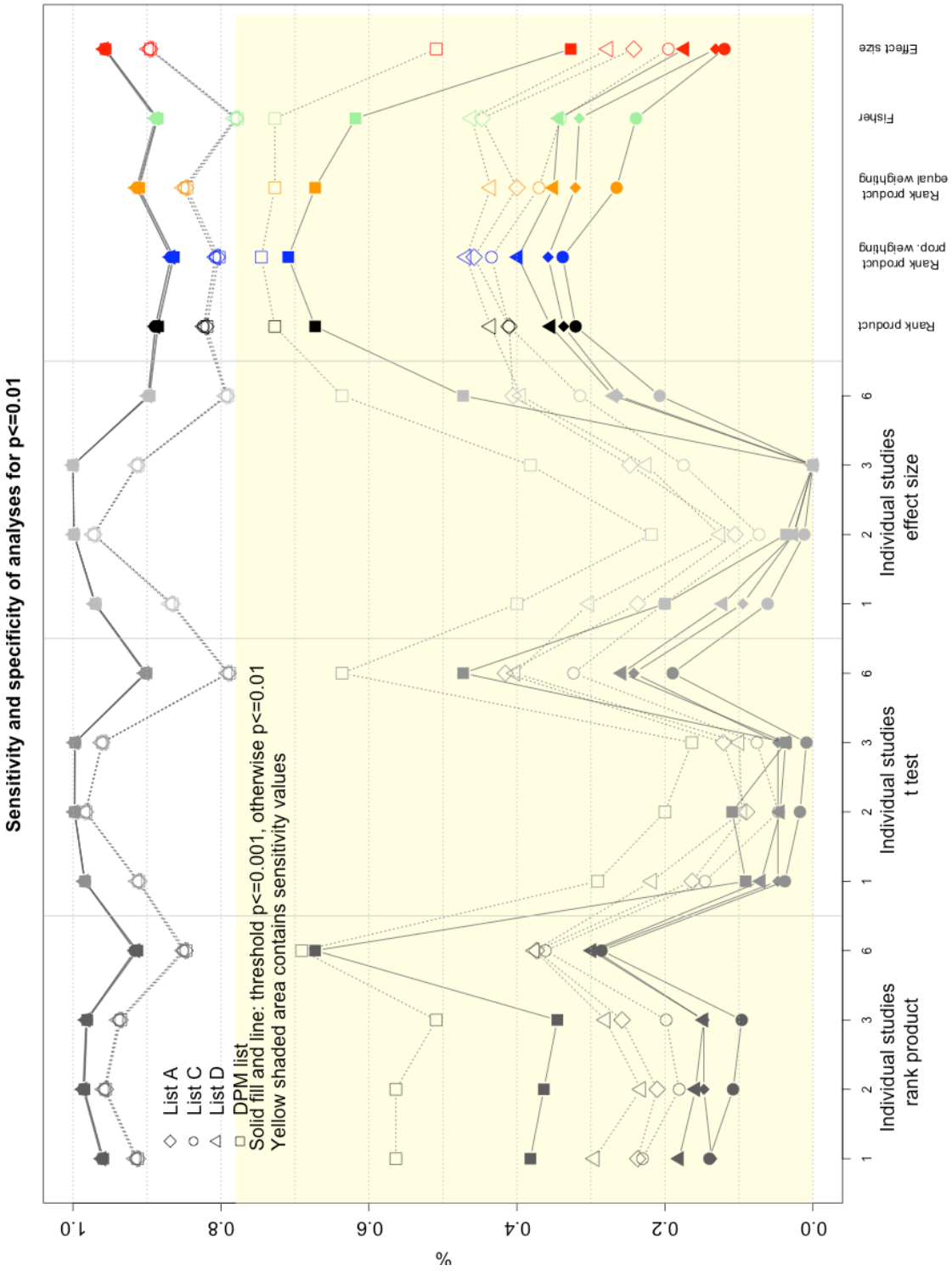
A4 – ROC plots for NCBI-derived gene reference lists

ROC plots for assessing overlap between meta-analysis results and computationally generated lists (A,C,D) and manually curated (DPM) interferon-related gene lists.



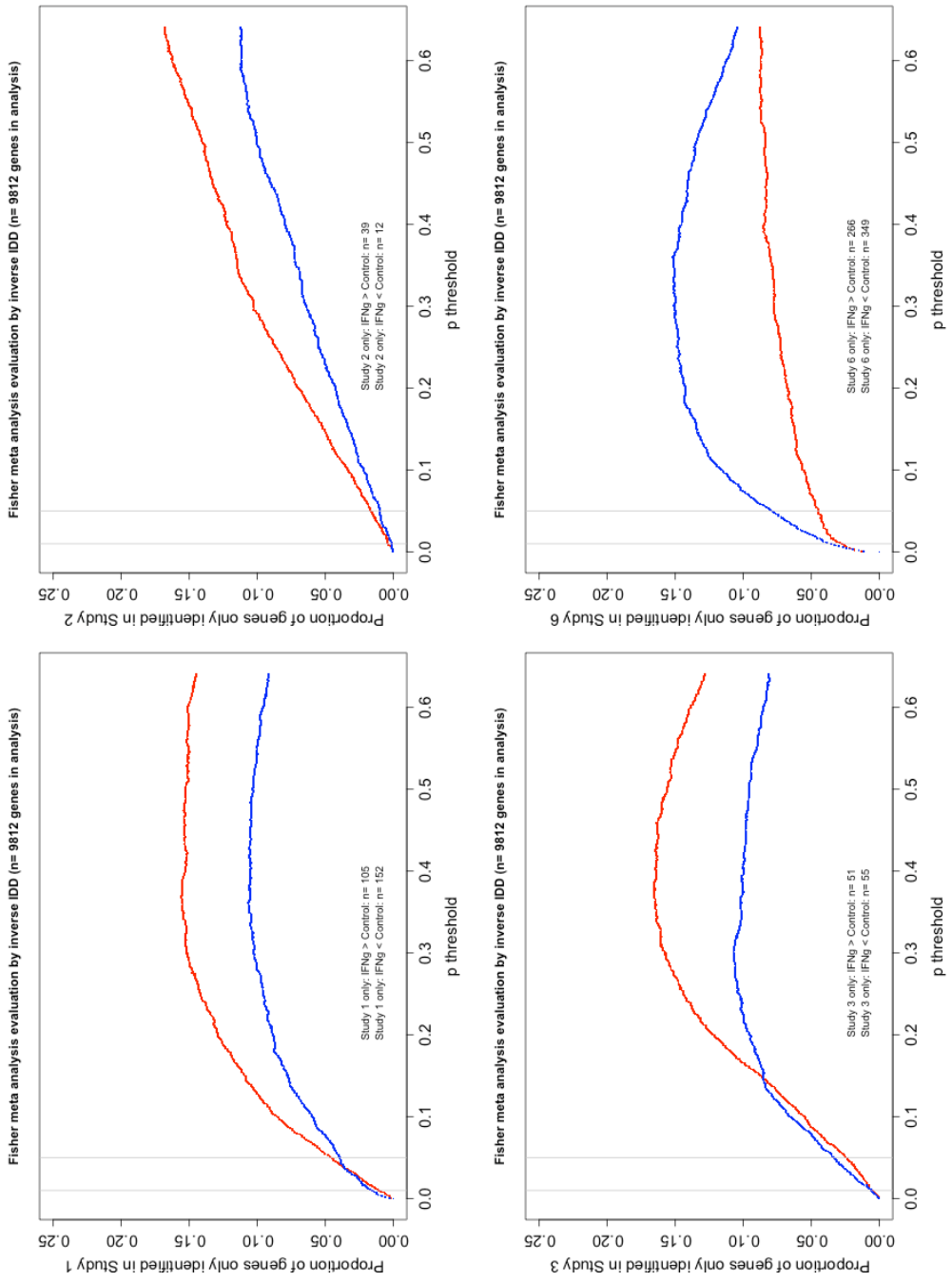
A5 – Sensitivity and specificity summary

Summary graph of sensitivity and specificity estimates when comparing overlap between statistical results (individual studies and meta-analysis, by any of the used methods, and for a given significance threshold) and biological reference lists A,C,D, and DPM gold standard list.



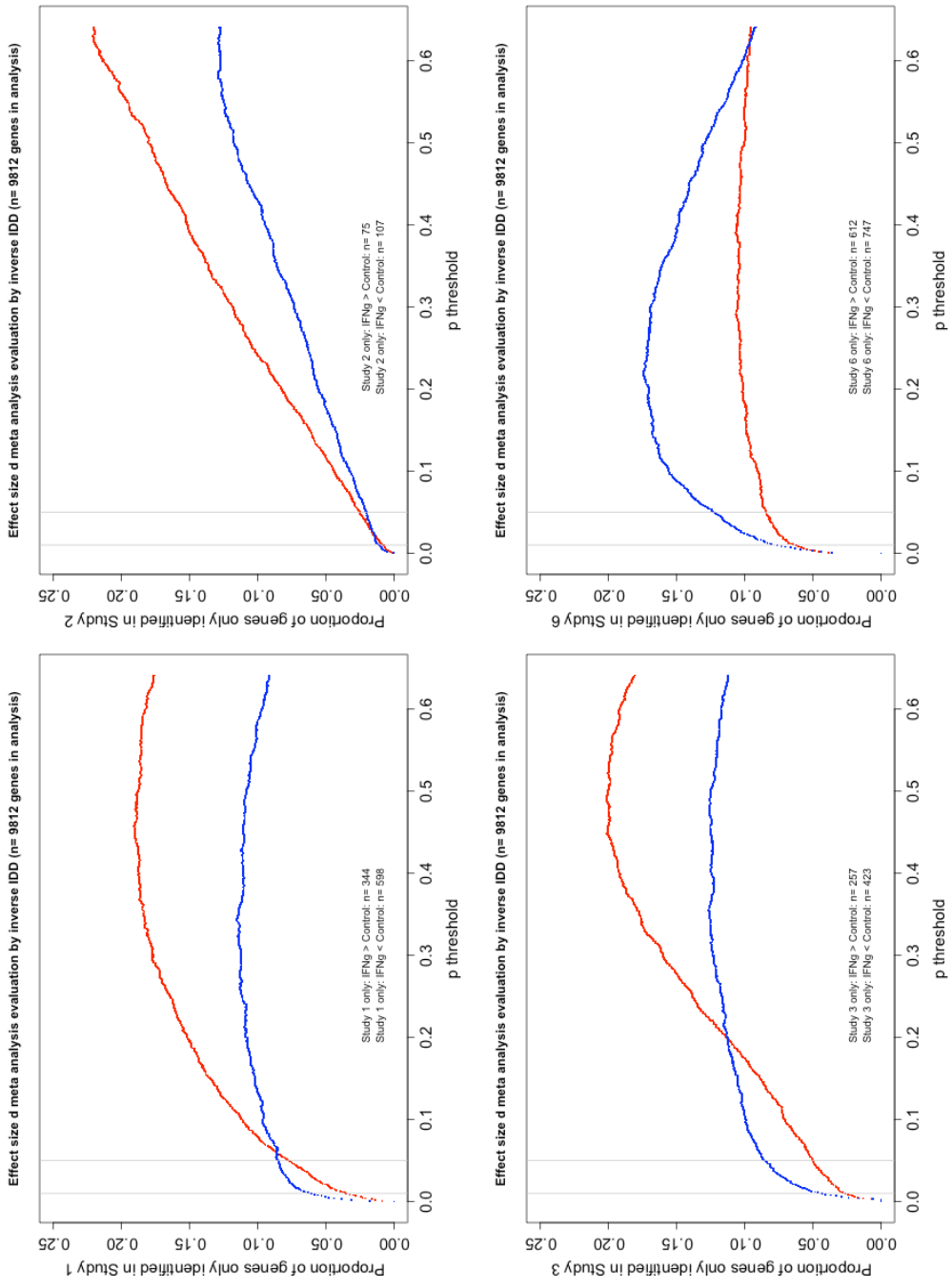
A6 – Reversed IDR plots, Fisher's meta-analysis

Independent Discovery Rate (IDR) plots reversed, i.e. number of statistical results unique to a study and not recovered by meta-analysis.



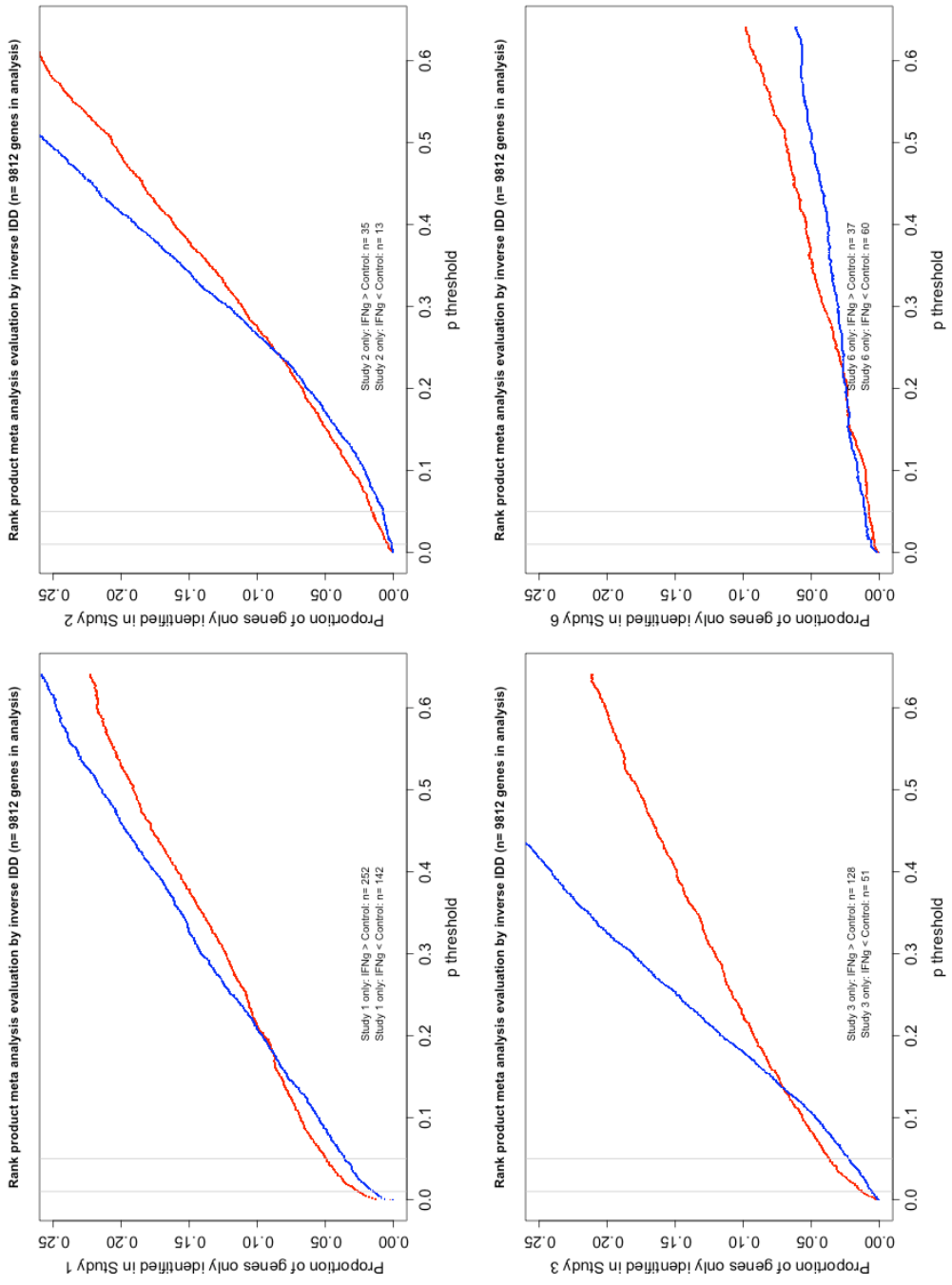
A7 – Reversed IDR plots, Effect Size meta-analysis

Independent Discovery Rate (IDR) plots reversed, i.e. number of statistical results unique to a study and not recovered by meta-analysis.



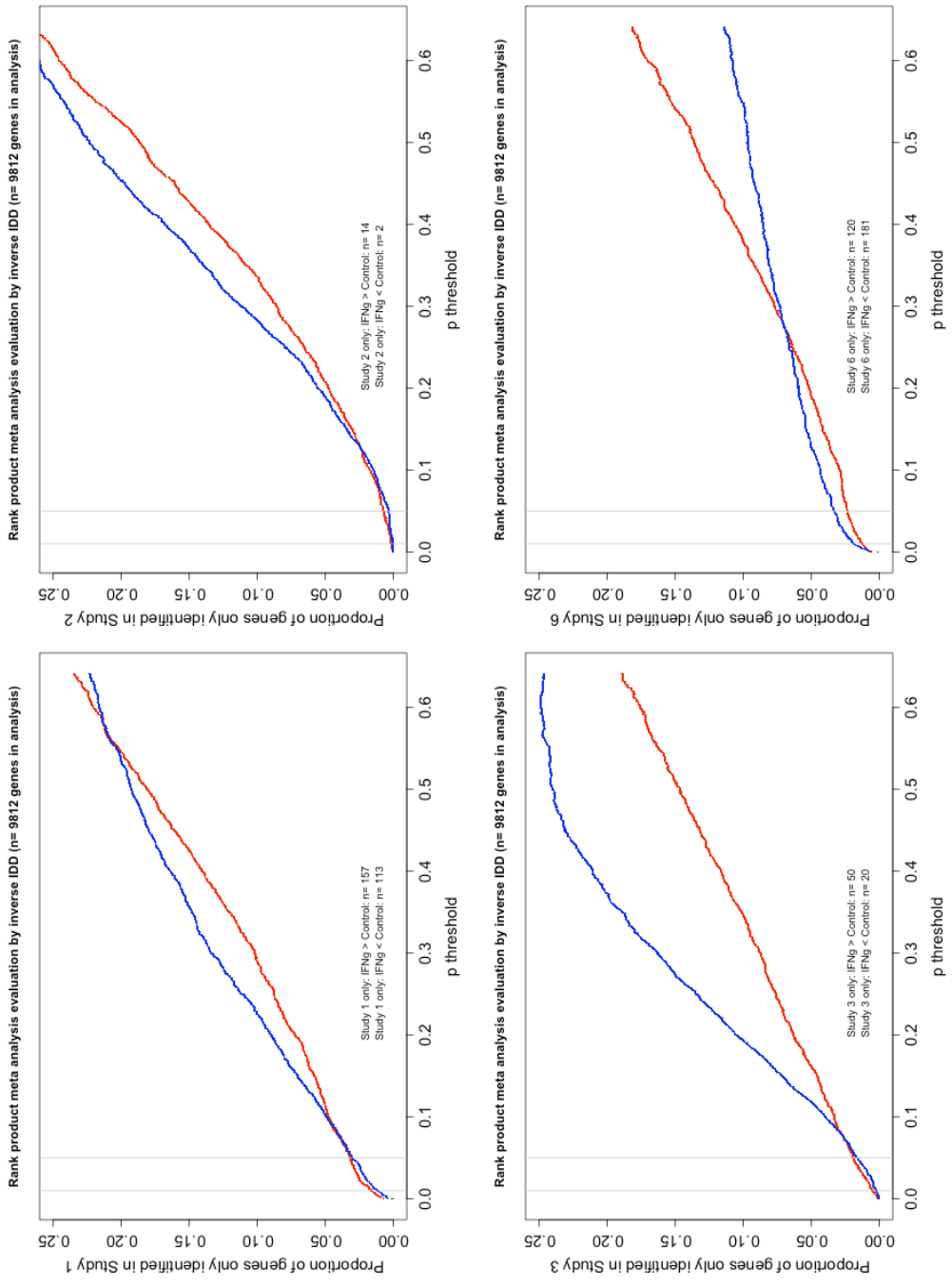
A8 – Reversed IDR plots, Rank Product meta-analysis

Independent Discovery Rate (IDR) plots reversed, i.e. number of statistical results unique to a study and not recovered by meta-analysis.



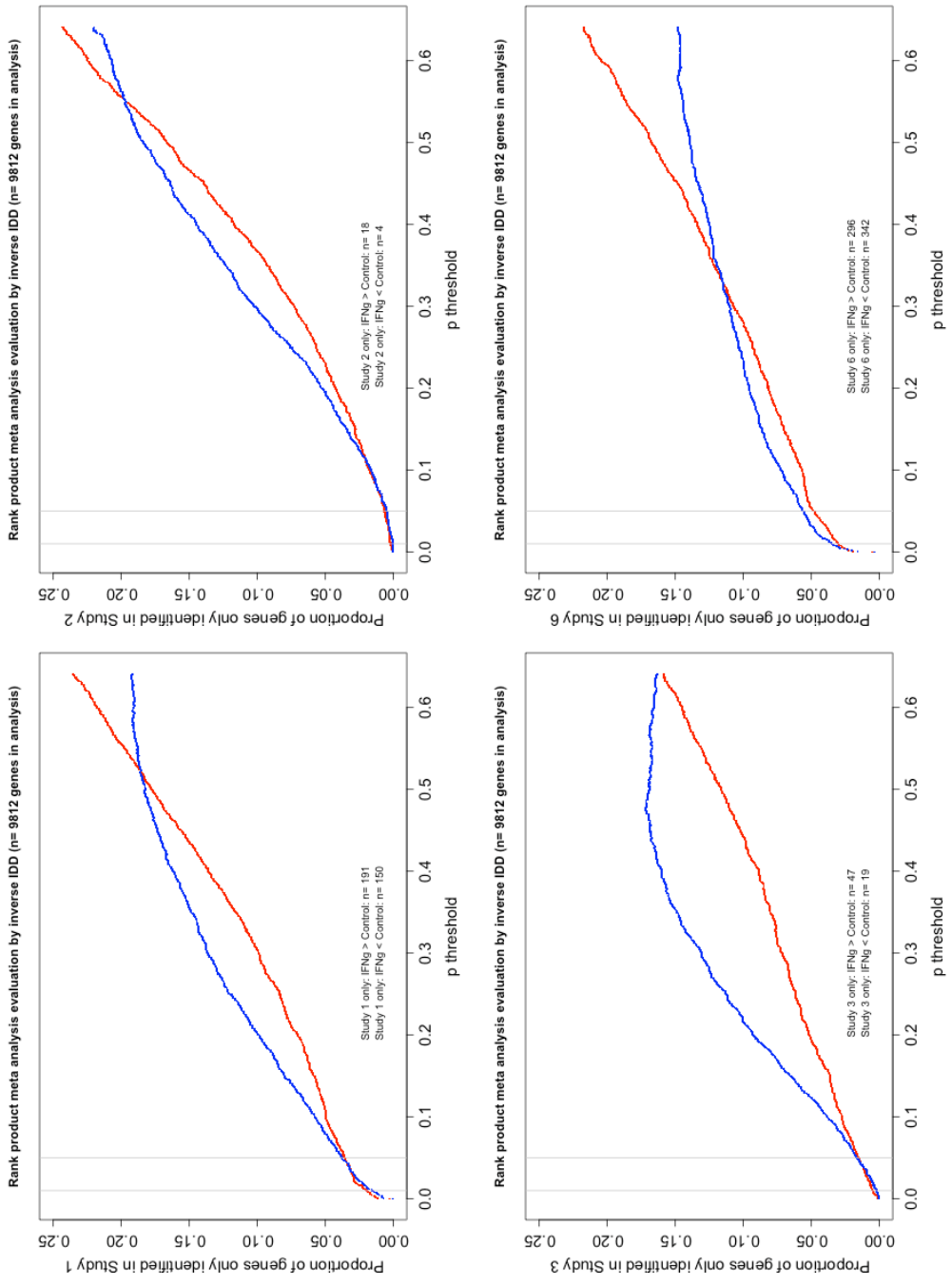
A9 – Reversed IDR plots, Rank Product meta-analysis, proportional weights

Independent Discovery Rate (IDR) plots reversed, i.e. number of statistical results unique to a study and not recovered by meta-analysis.



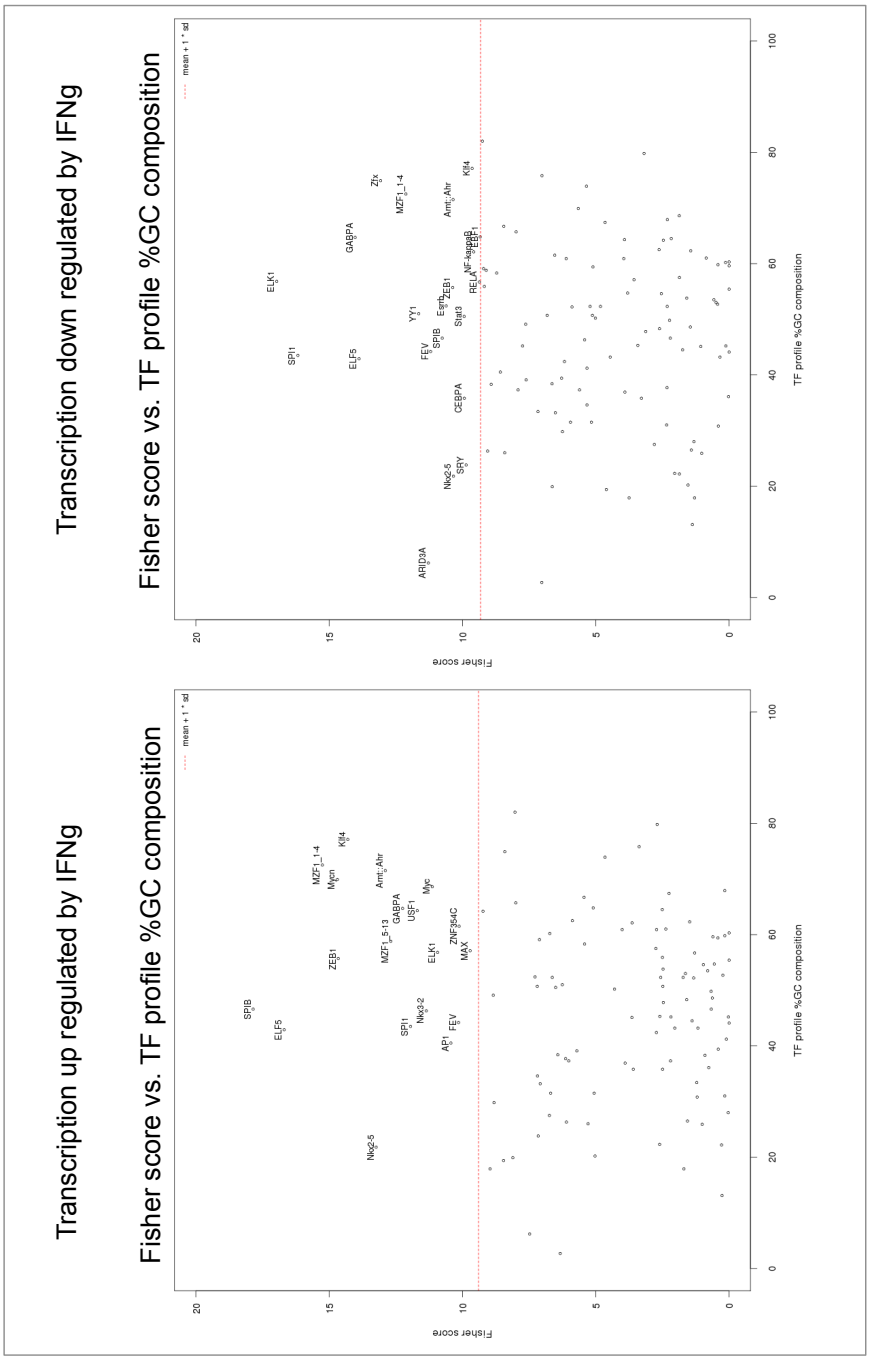
A10 – Reversed IDR plots, Rank Product meta-analysis, equal weights

Independent Discovery Rate (IDR) plots reversed, i.e. number of statistical results unique to a study and not recovered by meta-analysis.



A11 – Transcription factor enrichment of aggregated meta-analysis results

oPOSSUM enrichment analysis, testing for transcription factor binding site enrichment within meta-analysis results



A12 – DPM gene reference list ("gold standard")

This list has been manually curated by the Division of Pathway Medicine and the Ghazal group and reflects the core interferon gamma transcriptional network.

MGI Gene Symbol	Gene Description	Entrez Gene ID
Ciita	class II transactivator [Mus musculus]	12265
Ccl5	CHEMOKINE (C-C MOTIF) LIGAND 5	20304
Fasl	FAS LIGAND (TNF SUPERFAMILY, MEMBER 6)	14103
Gbp1	GUANYLATE NUCLEOTIDE BINDING PROTEIN 1	14468
H2-Aa	HISTOCOMPATIBILITY 2, CLASS II ANTIGEN A, ALPHA	14960
H2-Ab1	HISTOCOMPATIBILITY 2, CLASS II ANTIGEN A, BETA 1	14961
H2-Bl	HISTOCOMPATIBILITY 2, BLASTOCYST	14963
H2-DMa	HISTOCOMPATIBILITY 2, CLASS II, LOCUS DMA	14998
H2-DMb1	HISTOCOMPATIBILITY 2, CLASS II, LOCUS MB1	14999
H2-Ea	HISTOCOMPATIBILITY 2, CLASS II ANTIGEN E ALPHA	14968
H2-Eb1	HISTOCOMPATIBILITY 2, CLASS II ANTIGEN E BETA	14969
H2-M10.1	HISTOCOMPATIBILITY 2, M REGION LOCUS 10.1	14985
H2-M3	HISTOCOMPATIBILITY 2, M REGION LOCUS 3	14991
H2-M9	HISTOCOMPATIBILITY 2, M REGION LOCUS 9	14997
H2-Oa	HISTOCOMPATIBILITY 2, O REGION ALPHA LOCUS	15001
H2-Q1	HISTOCOMPATIBILITY 2, Q REGION LOCUS 1	15006
H2-Q10	HISTOCOMPATIBILITY 2, Q REGION LOCUS 10	15007
H2-Q2	HISTOCOMPATIBILITY 2, Q REGION LOCUS 2	15013
H2-Q5	HISTOCOMPATIBILITY 2, Q REGION LOCUS 5	15016
H2-Q7	HISTOCOMPATIBILITY 2, Q REGION LOCUS 7	15018
H2-T22	HISTOCOMPATIBILITY 2, T REGION LOCUS 22	15039
H2-T23	HISTOCOMPATIBILITY 2, T REGION LOCUS 23	15040
H2-T24	HISTOCOMPATIBILITY 2, T REGION LOCUS 24	15042
H2-T3	HISTOCOMPATIBILITY 2, T REGION LOCUS 3	15043
Il12b	INTERLEUKIN 12B	16160
Il15	INTERLEUKIN 15	16168
Irf1	INTERFERON REGULATORY FACTOR 1	16362
Irf2	INTERFERON REGULATORY FACTOR 2	16363
Psmb9	PROTEOSOME (PROSOME, MACROPAIN) SUBUNIT, BETA TYPE 9 (LARGE MULTIFUNCTIONAL PEPTIDASE 2)	16912
Socs1	SUPPRESSOR OF CYTOKINE SIGNALING 1	12703
Socs3	SUPPRESSOR OF CYTOKINE SIGNALING 3	12702
Stat1	SIGNAL TRANSDUCER AND ACTIVATOR OF TRANSCRIPTION 1	20846
Tap1	TRANSPORTER 1, ATP-BINDING CASSETTE, SUB-FAMILY B (MDR/TAP)	21354
Nfkbia	nuclear factor of kappa light polypeptide gene enhancer in B-cells inhibitor, alpha	18035
Tnf	tumor necrosis factor	21926
Tlr2	toll-like receptor 2	24088
CCl4	chemokine (C-C motif) ligand 4	20303
Gbp5	guanylate binding protein 5	
Cxcl10	chemokine (C-X-C motif) ligand 10	15945
Gbp3	guanylate binding protein 3	55932
Ifit1	interferon-induced protein with tetratricopeptide repeats 1	15957
Ifit3	interferon-induced protein with tetratricopeptide repeats 3	15959
Mx2	myxovirus (influenza virus) resistance 2	17858
Adar	adenosine deaminase, RNA-specific	56417

Stat2	signal transducer and activator of transcription 2	20847
Ddx58	DEAD (Asp-Glu-Ala-Asp) box polypeptide 58	
Ccl3	chemokine (C-C motif) ligand 3	20302
Cd274	CD274 antigen	60533
Oasl1	2'-5' oligoadenylate synthetase-like 1	231655
Oas2	2'-5' oligoadenylate synthetase-like 2	23962
Gbp2	guanylate binding protein 2	14469
Gbp6	guanylate binding protein 6	229900
Tlr3	toll-like receptor 3	142980
Ifih1	interferon induced with helicase C domain 1	71586
Ifit2	interferon-induced protein with tetratricopeptide repeats 2	15958
Ifi205	interferon activated gene 205	226695
Isg20	interferon-stimulated protein	57444
Zbp 1	Z-DNA binding protein 1	58203
Gbp4	guanylate binding protein 4	
H2-D1	histocompatibility 2, D region locus 1	14964
Irf3	interferon regulatory factor 3	54131
Irf4	interferon regulatory factor 4	16364
Irf5	interferon regulatory factor 5	27056
Irf6	interferon regulatory factor 6	54139
Irf7	interferon regulatory factor 7	54123
Irf8	interferon regulatory factor 8	15900
Irf9	interferon regulatory factor 9	16391

Appendix – Publications (co/authored)

Relevant to the contents of this thesis

The Transcription Factor STAT-1 Couples Macrophage Synthesis of 25-Hydroxycholesterol to the Interferon Antiviral Response
Blanc, M., Hsieh, W. Y., Robertson, K. A., Kropp, K. A., Forster, T., Shui, G., Lacaze, P., Watterson, S., Griffiths, S. J., Spann, N. J., Meljon, A., Talbot, S., Krishnan, K., Covey, D. F., Wenk, M. R., Craigon, M., Ruzsics, Z., Haas, J., Angulo, A., Griffiths, W. J., Glass, C. K., Wang, Y. & Ghazal, P.24-Jan-**2013** In : **Immunity**.38, (1); 106-18

Host defense against viral infection involves interferon mediated down-regulation of sterol biosynthesis
Blanc, M., Hsieh, W. Y., Robertson, K. A., Watterson, S., Shui, G., Lacaze, P., Khondoker, M., Dickinson, P., Sing, G., Rodríguez-Martín, S., Phelan, P., Forster, T., Strobl, B., Müller, M., Riemersma, R., Osborne, T., Wenk, M. R., Angulo, A. & Ghazal, P.**2011** In : **PLoS Biology**.9, 3, p. e1000598

Reversible inhibition of murine cytomegalovirus replication by gamma interferon (IFN-γ) in primary macrophages involves a primed type I IFN-signaling subnetwork for full establishment of an immediate-early antiviral state
Kropp, K. A., Robertson, K. A., Sing, G., Rodriguez-Martin, S., Blanc, M., Lacaze, P., Hassim, M. F. B. N., Khondoker, M. R., Busche, A., Dickinson, P., Forster, T., Strobl, B., Mueller, M., Jonjic, S., Angulo, A. & Ghazal, P.**2011** In : **Journal of Virology**.85, 19, p. 10286-99, 14 p.

Combined genome-wide expression profiling and targeted RNA interference in primary mouse macrophages reveals perturbation of transcriptional networks associated with interferon signalling
Lacaze, P., Raza, S., Sing, G., Page, D., Forster, T., Storm, P., Craigon, M., Awad, T., Ghazal, P. & Freeman, T. C.Aug-**2009** In : **BMC Genomics**.10, 372

Estimation of expression levels in spotted microarrays with saturated pixels
Glasbey, C. A., Forster, T. & Ghazal, P.**2007** In : **Statistical applications in genetics and molecular biology**.6, p. Article34

Modelling of macrophage gene expression in the interferon pathway
Yu, L., Marshall, S., Forster, T. & Ghazal, P.**2006** *2006 IEEE International Workshop on Genomic Signal Processing and Statistics.* NEW YORK: IEEE, p. 45-462 p.

GPX-Macrophage Expression Atlas: a database for expression profiles of macrophages challenged with a variety of pro-inflammatory, anti-inflammatory, benign and pathogen insults
Grimes, G. R., Moodie, S., Beattie, J. S., Craigon, M., Dickinson, P., Forster, T., Livingston, A. D., Mewissen, M., Robertson, K. A., Ross, A. J., Sing, G. & Ghazal, P.12-Dec-2005 In : BMC Genomics.6, 178, p. 1-8, 8 p.

Microarrays - Standard operating procedures.
Ghazal, P. & Forster, T.2005*Encyclopedia of Medical Genomics and Proteomics.* Fuchs, J. (ed.). MARCEL DEKKER, p. 824-8306 p.

Other

Exploiting Parallel R in the Cloud with SPRINT
Piotrowski, M., McGilvary, G. A., Sloan, T. M., Mewissen, M., Lloyd, A. D., Forster, T., Mitchell, L., Ghazal, P. & Hill, J. Jan-2013 In : Methods of Information in Medicine.52, 1, p. 80-90, 11 p.

SPRINT v1.0.4 Software Release
Troup, E., Mitchell, L., Sloan, T., Robertson, K., Forster, T. & Ghazal, P.Jan-2013

Parallel classification and feature selection in microarray data using SPRINT
Mitchell, L., Sloan, T. M., Mewissen, M., Ghazal, P., Forster, T., Piotrowski, M. & Trew, A.2012 In : Concurrency and Computation: Practice and Experience.p. n/a-n/a

Optimization of a parallel permutation testing function for the SPRINT R package
Petrou, S., Sloan, T., Mewissen, M., Forster, T., Piotrowski, M., Dobrzelecki, B., Ghazal, P., Trew, A. & Hill, J. 10-Dec-2011 In : Concurrency and Computation: Practice and Experience.23, 17, p. 2258-2268, 11 p.

SPRINT: Parallel computing with R on HECToR: HECToR Training Course presented at NAG, Oxford, 1st Dec 2011
Piotrowski, M., Mewissen, M., Sloan, T., Forster, T., Mitchell, L. & Ghazal, P.1-Dec-2011

Proteolysis-inducing factor core peptide mediates dermcidin-induced proliferation of hepatic cells through multiple signalling networks
Lowrie, A. G., Dickinson, P., Wheelhouse, N., Stewart, G. D., Ross, A. J., Forster, T. & Ross, J. A.Sep-2011 In : International journal of oncology.39, 3, p. 709-718, 10 p.

Optimisation and parallelisation of the partitioning around medoids function in R
Piotrowski, M., Forster, T., Dobrezelecki, B., Sloan, T., Mitchell, L., Ghazal, P., Mewsissen, M., Petrou, S., Trew, A. & Hill, J. 1-Jul-2011p. 707-7137 p.

A parallel random forest classifier for R
Mitchell, L., Sloan, T., Mewissen, M., Ghazal, P., Forster, T., Piotrowski, M. & Trew, A. S.1-Jan-2011p. 1-67 p.

Optimization of a parallel permutation testing function for the SPRINT R package
Petrou, S., Sloan, T. M., Mewissen, M., Forster, T., Piotrowski, M., Dobrzelecki, B., Ghazal, P., Trew, A. & Hill, J. 2011 In : Concurrency and Computation: Practice and Experience.23, 17, p. 2258-2268, 11 p.

Temporal profiling of the coding and noncoding murine cytomegalovirus transcriptomes
Lacaze, P., Forster, T., Ross, A., Kerr, L. E., Salvo-Chirnside, E., Lisnic, V. J., López-Campos, G. H., García-Ramírez, J. J., Messerle, M., Trgovcich, J., Angulo, A. & Ghazal, P.2011 In : Journal of Virology.85, 12, p. 6065-76, 12 p.

Assessment of transcriptomal analysis of varicella-zoster-virus gene expression in patients with and without post-herpetic neuralgia
Ashrafi, G. H., Grinfeld, E., Montague, P., Forster, T., Ross, A., Ghazal, P., Scott, F., Breuer, J.,

Goodwin, R. & Kennedy, P. G. E. Oct-2010 In : Virus Genes.41, 2, p. 192-201, 10 p.

Combined agonist-antagonist genome-wide functional screening identifies broadly active antiviral microRNAs

Santhakumar, D., Forster, T., Laqtom, N. N., Fragkoudis, R., Dickinson, P., Abreu-Goodger, C., Manakov, S. A., Choudhury, N. R., Griffiths, S. J., Vermeulen, A., Enright, A. J., Dutia, B., Kohl, A., Ghazal, P. & Buck, A. H.Aug-2010 In : Proceedings of the National Academy of Sciences of the United States of America - PNAS.107, 31, p. 13830-13835, 6 p.

SPRINT: a Simple Parallel INTerface to High Performance Computing and a Parallel R Function Library.
Mewissen, M., Forster, T., Sloan, T., Petrou, S., Piotrowski, M., Dobrzelecki, B., Ghazal, P., Trew, A. & Hill, J. Jul-2010p. 1051 p.

Quantitative analysis of low-abundance peptides in HeLa cell cytoplasm by targeted liquid chromatography/mass spectrometry and stable isotope dilution: emphasising the distinction between peptide detection and peptide identification
Le Bihan, T., Grima, R., Martin, S., Forster, T. & Le Bihan, Y. Apr-2010 In : Rapid communications in mass spectrometry.24, 7, p. 1093-1104, 12 p.

Applying statistical inference in genomics with evidence-based pathways: Towards elucidating new functional correlations of biomarkers.
Ghazal, P., Forster, T. & Ivens, A.22-Feb-2010 In : European Pharmaceutical Review.1, p. 12-15

Optimization of a parallel permutation testing function for the SPRINT R package
Petrou, S., Sloan, T., Mewissen, M., Forster, T., Piotrowski, M. & Dobrzelecki, B. 1-Jan-2010p. 516-5216 p.

Statistical methods for analysis of high-throughput RNA interference screens
Birmingham, A., Selfors, L. M., Forster, T., Wrobel, D., Kennedy, C. J., Shanks, E., Santoyo-Lopez, J., Dunican, D. J., Long, A., Kelleher, D., Smith, Q., Beijersbergen, R. L., Ghazal, P. & Shamu, C. E. Aug-2009 In : Nature Methods.6, 8, p. 569-575, 7 p.

Genome-wide reduction in transcriptomal profiles of varicella-zoster virus vaccine strains compared with Parental Oka strain using long oligonucleotide microarrays
Grinfeld, E., Ross, A., Forster, T., Ghazal, P. & Kennedy, P. Feb-2009 In : Virus Genes.38, 1, p. 19-29, 11 p.

Transcriptomal analysis of Varicella-Zoster-Virus gene expression in patients with and without Post-Herpetic Neuralgia
Ashrafi, H., Grinfeld, E., Forster, T., Ross, A., Ghazal, P., Breuer, J., Goodwin, R. & Kennedy, P. G. E. 2009 In : Journal of Neurovirology.15, p. 77, 1 p.

High-resolution gene expression profiling for simultaneous kinetic parameter analysis of RNA synthesis and decay
Dolken, L., Ruzsics, Z., Radle, B., Friedel, C. C., Zimmer, R., Mages, J., Hoffmann, R., Dickinson, P., Forster, T., Ghazal, P. & Koszinowski, U. H. Sep-2008 In : RNA.14, 9, p. 1959-1972, 14 p.

Evaluation of a protein microarray method for immuno-typing erythrocytes in whole blood
O'Looney, N., Burgess, S. T. G., Kwan, M. C., Ross, A. J., Robb, J., Forster, T., Beattie, J. S., Ghazal, P., Petrik, J. & Campbell, C. J. 2008 In : Journal of Immunoassay and Immunochemistry.29, 2, p. 197-209, 13 p.

Molecular profiling of the human testis reveals stringent pathway-specific regulation of RNA expression following gonadotropin suppression and progestogen treatment
Bayne, R. A. L., Forster, T., Burgess, S. T. G., Craigon, M., Walton, M. J., Baird, D. T., Ghazal, P. & Anderson, R. A.2008 In : Journal of Andrology.29, 4, p. 389-403, 15 p.

SPRINT: a new parallel framework for R
Hill, J., Hambley, M., Forster, T., Mewissen, M., Sloan, T., Scharinger, F., Trew, A. & Ghazal, P.2008 In : BMC Bioinformatics.9, p. 558

Investigation of host RNA responses to infection in neonates: a new avenue for diagnosis?
Smith, C. L., Dickinson, P., Forster, T., Lacaze, P., Stenson, B. J. & Ghazal, P.Oct-2007 In : Acta paediatrica.96, p. 88, 1 p.

Quantitative assessment of human whole blood RNA as a potential biomarker for infectious disease
Smith, C. L., Dickinson, P., Forster, T., Khondoker, M., Craigon, M., Ross, A., Storm, P., Burgess, S., Lacaze, P., Stenson, B. J. & Ghazal, P.2007 In : Analyst.132, 12, p. 1200-9, 10 p.

Innate immune response gene expression profiles of N9 microglia are pathogen-type specific
McKimmie, C. S., Roy, D., Forster, T. & Fazakerley, J. K. Jun-2006 In : Journal of neuroimmunology.175, 1-2, p. 128-141, 14 p.

Gene expression profiling of mid to late secretory phase endometrial biopsies from women with menstrual complaint
Critchley, H. O. D., Robertson, K. A., Forster, T., Henderson, T. A., Williams, A. R. W. & Ghazal, P.2006 In : American journal of obstetrics and gynecology.195, 2, p. 406.e1-16

Locus-specific gene expression pattern suggests a unique propagation strategy for a giant algal virus
Allen, M. J., Forster, T., Schroeder, D. C., Hall, M., Roy, D., Ghazal, P. & Wilson, W. H. 2006 In : Journal of Virology.80, 15, p. 7699-705, 7 p.

Microarray and bioarray technology.
Ghazal, P. & Forster, T.2005*Bioinformatics Baiscs: Applications in biological science and medicine.: Second Edicition.* Buehler, L. K. & Rashidi, H. H. (eds.). CRC press, p. 190-21626 p.

Transcriptomal analysis of varicella-zoster virus infection using long oligonucleotide-based microarrays
Kennedy, P. G. E., Grinfeld, E., Craigon, M., Vierlinger, K., Roy, D., Forster, T. & Ghazal, P.2005 In : Journal of General Virology.86, Pt 10, p. 2673-84, 12 p.

Triple-target microarray experiments: a novel experimental strategy
Forster, T., Costa, Y., Roy, D., Cooke, H. J. & Maratou, K. 10-Feb-2004 In : BMC Genomics.5, p. -, 16 p., 13

Expression profiling of the developing testis in wild-type and Dazl knockout mice
Maratou, K., Forster, T., Costa, Y., Taggart, M., Speed, R. M., Ireland, J., Teague, P., Roy, D. & Cooke, H. J. Jan-2004 In : Molecular Reproduction and Development.67, 1, p. 26-54, 29 p.

Steroid signalling in human ovarian surface epithelial cells: the response to interleukin-1alpha determined by microarray analysis
Rae, M. T., Niven, D., Ross, A., Forster, T., Lathe, R., Critchley, H. O. D., Ghazal, P. & Hillier, S. G. 2004 In : Journal of Endocrinology.183, 1, p. 19-28, 10 p.

Gene expression responses induced in kidney by aldosterone excess and salt
Marshall, E., Dickinson, P., Forster, T., Roy, D., Kenyon, C. J. & Brown, R. W. Nov-2003 In : Journal of the American Society of Nephrology.14, p. 140A-140A, 1 p.

Experiments using microarray technology: limitations and standard operating procedures
Forster, T., Roy, D. & Ghazal, P.2003 In : Journal of Endocrinology.178, 2, p. 195-204, 10 p.

Meeting review: Bioinformatics of biochips: Accelerating discovery in functional genomics
Roy, D., Forster, T., McGeever, S., Robertson, K. N. & Ghazal, P.Aug-2002 In : Comparative and functional genomics.3, 4, p. 380-386, 7 p.

Outcome after extended follow-up in a prospective study of operable breast cancer: key factors and a prognostic index
Hawkins, R. A., Tesdale, A. L., Prescott, R. J., Forster, T., McIntyre, M. A., Baker, P., Jack, W. J. L., Chetty, U., Dixon, J. M., Killen, M. E., Hulme, M. J. & Miller, W. R. 1-Jul-2002 In : British Journal of Cancer.87, 1, p. 814, 7 p.

Effect of a national community intervention programme on healing rates of chronic leg ulcer: Randomised controlled trial
Brown, A., Burns, E., Chalmers, L., Corcoran, F., Dale, J., Douglas, S., Finnie, A., Forster, T., Fowkes, G., Gibson, B., Gillies, T., Hill, E., Lee, R., MacGuire, C., Morley, S., O'Hare, J., Prescott, R., Ruckley, V., Young, C. & Scottish Leg Ulser Trial Participa 2002 In : Phlebology.17, 2, p. 47-53, 7 p.

Bibliography

Aderem, A. and D. M. Underhill (1999). "Mechanisms of phagocytosis in macrophages." Annual Review of Immunology **17**: 593-623.

Allen, L. A. H., C. M. Yang and J. E. Pessin (2002). "Rate and extent of phagocytosis in macrophages lacking vamp3." Journal of Leukocyte Biology **72**(1): 217-221.

Anders, S. and W. Huber (2010). "Differential expression analysis for sequence count data." Genome Biol **11**(10): R106.

Anderson, M. K., A. H. Weiss, G. Hernandez-Hoyos, C. J. Dionne and E. V. Rothenberg (2002). "Constitutive expression of PU.1 in fetal hematopoietic progenitors blocks T cell development at the pro-T cell stage." Immunity **16**(2): 285-296.

Angulo, A., P. Ghazal and M. Messerle (2000). "The major immediate-early gene ie3 of mouse cytomegalovirus is essential for viral growth." J Virol **74**(23): 11129-11136.

Angulo, A., P. Ghazal and M. Messerle (2000). "The major immediate-early gene ie3 of mouse cytomegalovirus is essential for viral growth." Journal of Virology **74**(23): 11129-11136.

Aschoff, I.-T. v. (1924). "Das reticulo-endotheliale system." Ergebnisse der inneren Medizin und Kinderheilkunde **26**: 1-118.

Bach, E. A., M. Aguet and R. D. Schreiber (1997). "The IFN gamma receptor: A paradigm for cytokine receptor signaling." Annual Review of Immunology **15**: 563-&.

Baron, D., A. Bihouee, R. Teusan, E. Dubois, F. Savagner, M. Steenman, R. Houlgatte and G. Ramstein (2011). "MADGene: retrieval and processing of gene identifier lists for the analysis of heterogeneous microarray datasets." Bioinformatics **27**(5): 725-726.

Battke, F. and K. Nieselt (2011). "Mayday SeaSight: combined analysis of deep sequencing and microarray data." PLoS One **6**(1): e16345.

Behrens, E. M., P. Gadue, S. Y. Gong, S. Garrett, P. L. Stein and P. L. Cohen (2003). "The mer receptor tyrosine kinase: expression and function suggest a role in innate immunity." European Journal of Immunology **33**(8): 2160-2167.

Beissbarth, T. and T. P. Speed (2004). "GOstat: find statistically overrepresented Gene Ontologies within a group of genes." Bioinformatics **20**(9): 1464-1465.

Benedict, C., M. Geisler, J. Trygg, N. Huner and V. Hurry (2006). "Consensus by democracy. Using meta-analyses of microarray and genomic data to model the cold acclimation signaling pathway in Arabidopsis." Plant Physiol **141**(4): 1219-1232.

Beyer, M., M. R. Mallmann, J. Xue, A. Staratschek-Jox, D. Vorholt, W. Krebs, D. Sommer, J. Sander, C. Mertens, A. Nino-Castro, S. V. Schmidt and J. L. Schultze (2012). "High-resolution transcriptome of human macrophages." PLoS One **7**(9): e45466.

Biesecker, L. G., W. Burke, I. Kohane, S. E. Plon and R. Zimmern (2012). "Next-generation sequencing in the clinic: are we ready?" Nature Reviews Genetics **13**(11): 818-824.

Billharz, R., H. Zeng, S. C. Proll, M. J. Korth, S. Lederer, R. Albrecht, A. G. Goodman, E. Rosenzweig, T. M. Tumpey, A. Garcia-Sastre and M. G. Katze (2009). "The NS1 Protein of the 1918 Pandemic Influenza Virus Blocks Host Interferon and Lipid Metabolism Pathways." Journal of Virology **83**(20): 10557-10570.

Biron, C. A. (1999). "Initial and innate responses to viral infections - pattern setting in immunity or disease." Current Opinion in Microbiology **2**(4): 374-381.

Blanc, M., W. Y. Hsieh, K. A. Robertson, S. Watterson, G. H. Shui, P. Lacaze, M. Khondoker, P. Dickinson, G. Sing, S. Rodriguez-Martin, P. Phelan, T. Forster, B. Strobl, M. Muller, R. Riemersma, T. Osborne, M. R. Wenk, A. Angulo and P. Ghazal (2011). "Host Defense

against Viral Infection Involves Interferon Mediated Down-Regulation of Sterol Biosynthesis." Plos Biology **9**(3).

Blanc, M., Hsieh, WY, Robertson, KA, Kropp, KA, Forster, T, Shui, G, Lacaze, P, Watterson, S, Griffiths, SJ, Spann, NJ, Meljon, A, Talbot, S, Krishnan, K, Covey, DF, Wenk, MR, Craigon, M, Ruzsics, Z, Haas, J, Angulo, A, Griffiths, WJ, Glass, CK, Wang, Y & Ghazal, P (2013). "The Transcription Factor STAT-1 Couples Macrophage Synthesis of 25-Hydroxycholesterol to the Interferon Antiviral Response." Immunity **38**(1): 13.

Blanchette, J., M. Jaramillo and M. Olivier (2003). "Signalling events involved in interferon-gamma-inducible macrophage nitric oxide generation." Immunology **108**(4): 513-522.

Bo, T. H., J. Dysvik and I. Jonassen (2004). "LSimpute: accurate estimation of missing values in microarray data with least squares methods." Nucleic Acids Research **32**(3).

Boehm, U., T. Klamp, M. Groot and J. C. Howard (1997). "Cellular responses to interferon-gamma." Annual Review of Immunology **15**: 749-795.

Bohn, E., J. Heesemann, S. Ehlers and I. B. Autenrieth (1994). "Early gamma interferon mRNA expression is associated with resistance of mice against Yersinia enterocolitica." Infect Immun **62**(7): 3027-3032.

Bollinger, T., A. Leutz, A. Leliavski, L. Skrum, J. Kovac, L. Bonacina, C. Benedict, T. Lange, J. Westermann, H. Oster and W. Solbach (2011). "Circadian Clocks in Mouse and Human CD4+T Cells." Plos One **6**(12).

Bolstad, B. "preprocessCore: A collection of pre-processing functions."

Bolstad, B. M., R. A. Irizarry, M. Astrand and T. P. Speed (2003). "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias." Bioinformatics **19**(2): 185-193.

Borthwick, F., A. M. Allen, J. M. Taylor and A. Graham (2010). "Overexpression of STARD3 in human monocyte/macrophages induces an anti-atherogenic lipid phenotype." Clinical Science **119**(7-8): 265-272.

Brancho, D., N. Tanaka, A. Jaeschke, J. J. Ventura, N. Kelkar, Y. Tanaka, M. Kyuuma, T. Takeshita, R. A. Flavell and R. J. Davis (2003). "Mechanism of p38 MAP kinase activation in vivo." Genes & Development **17**(16): 1969-1978.

Brazma, A., P. Hingamp, J. Quackenbush, G. Sherlock, P. Spellman, C. Stoeckert, J. Aach, W. Ansorge, C. A. Ball, H. C. Causton, T. Gaasterland, P. Glenisson, F. C. P. Holstege, I. F. Kim, V. Markowitz, J. C. Matese, H. Parkinson, A. Robinson, U. Sarkans, S. Schulze-Kremer, J. Stewart, R. Taylor, J. Vilo and M. Vingron (2001). "Minimum information about a microarray experiment (MIAME) - toward standards for microarray data." Nature Genetics **29**(4): 365-371.

Breiman, A., N. Grandvaux, R. T. Lin, C. Ottone, S. Akira, M. Yoneyama, T. Fujita, J. Hiscott and E. F. Meurs (2005). "Inhibition of RIG-I-dependent signaling to the interferon pathway during hepatitis C virus expression and restoration of signaling by IKK epsilon." Journal of Virology **79**(7): 3969-3978.

Breitling, R., P. Armengaud, A. Amtmann and P. Herzyk (2004). "Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments." Febs Letters **573**(1-3): 83-92.

Brien, S. E., P. E. Ronksley, B. J. Turner, K. J. Mukamal and W. A. Ghali (2011). "Effect of alcohol consumption on biological markers associated with risk of coronary heart disease: systematic review and meta-analysis of interventional studies." BMJ **342**: d636.

Brock, G. N., J. R. Shaffer, R. E. Blakesley, M. J. Lotz and G. C. Tseng (2008). "Which missing value imputation method to use in expression profiles: a comparative study and two selection schemes." Bmc Bioinformatics **9**.

Brodeur, J., C. Theriault, M. Lessard-Beaudoin, A. Marcil, S. Dahan and C. Lavoie (2012). "LDLR-related protein 10 (LRP10) regulates amyloid precursor protein (APP) trafficking and processing: evidence for a role in Alzheimer's disease." Molecular Neurodegeneration **7**.

Busby, M. A., C. Stewart, C. A. Miller, K. R. Grzeda and G. T. Marth (2013). "Scotty: a web tool for designing RNA-Seq experiments to measure differential gene expression." Bioinformatics **29**(5): 656-657.

Cahan, P., F. Rovegno, D. Mooney, J. C. Newman, G. St Laurent, 3rd and T. A. McCaffrey (2007). "Meta-analysis of microarray results: challenges, opportunities, and recommendations for standardization." Gene **401**(1-2): 12-18.

Camenisch, T. D., B. H. Koller, H. S. Earp and G. K. Matsushima (1999). "A novel receptor tyrosine kinase, Mer, inhibits TNF-alpha production and lipopolysaccharide-induced endotoxic shock." Journal of Immunology **162**(6): 3498-3503.

Carstensen, B. P., M; Laara, E; Hills, M (2012). "Epi: A Package for Statistical Analysis in Epidemiology."

Cavadini, G., S. Petrzilka, P. Kohler, C. Jud, I. Tobler, T. Birchler and A. Fontana (2007). "TNF-alpha suppresses the expression of clock genes by interfering with E-box-mediated transcription." Proceedings of the National Academy of Sciences of the United States of America **104**(31): 12843-12848.

Chalmers, I., L. V. Hedges and H. Cooper (2002). "A brief history of research synthesis." Evaluation & the Health Professions **25**(1): 12-37.

Chavan, S. S., M. A. Bauer, E. A. Peterson, C. J. Heuck and D. J. Johann (2013). "Towards the integration, annotation and association of historical microarray experiments with RNA-seq." BMC Bioinformatics **14 Suppl 14**: S4.

Choi, H., R. Shen, A. M. Chinnaiyan and D. Ghosh (2007). "A latent variable approach for meta-analysis of gene expression data from multiple microarray experiments." Bmc Bioinformatics **8**.

Choi, J. K., U. Yu, S. Kim and O. J. Yoo (2003). "Combining multiple microarray studies and modeling interstudy variation." Bioinformatics **19**: i84-i90.

Churchill, G. A. (2002). "Fundamentals of experimental design for cDNA microarrays." Nature Genetics **32 Suppl**: 490-495.

Cline, M. S., M. Smoot, E. Cerami, A. Kuchinsky, N. Landys, C. Workman, R. Christmas, I. Avila-Campilo, M. Creech, B. Gross, K. Hanspers, R. Isserlin, R. Kelley, S. Killcoyne, S. Lotia, S. Maere, J. Morris, K. Ono, V. Pavlovic, A. R. Pico, A. Vailaya, P. L. Wang, A. Adler, B. R. Conklin, L. Hood, M. Kuiper, C. Sander, I. Schmulevich, B. Schwikowski, G. J. Warner, T. Ideker and G. D. Bader (2007). "Integration of biological networks and gene expression data using Cytoscape." Nature Protocols **2**(10): 2366-2382.

Cochran, W. G. (1954). "The Combination of Estimates from Different Experiments." Biometrics **10**(1): 101-129.

Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences, Lawrence Erlbaum Associates.

Cox, D., D. J. Lee, B. M. Dale, J. Calafat and S. Greenberg (2000). "A Rab11-containing rapidly recycling compartment in macrophages that promotes phagocytosis." Proceedings of the National Academy of Sciences of the United States of America **97**(2): 680-685.

Crawford, R. M., D. S. Finbloom, J. Ohara, W. E. Paul and M. S. Meltzer (1987). "B cell stimulatory factor-1 (interleukin 4) activates macrophages for increased tumoricidal activity and expression of Ia antigens." Journal of Immunology **139**(1): 135-141.

Dalma-Weiszhausz, D. D., J. Warrington, E. Y. Tanimoto and G. Miyada (2006). "The affymetrix GeneChip (R) platform: An overview." DNA Microarrays Part A: Array Platforms and Wet-Bench Protocols **410**: 3-28.

Danilkovitch-Miagkova, A., F. M. Duh, I. Kuzmin, D. Angeloni, S. L. Liu, A. D. Miller and M. I. Lerman (2003). "Hyaluronidase 2 negatively regulates RON receptor tyrosine kinase and mediates transformation of epithelial cells by jaagsiekte sheep retrovirus." Proceedings of the National Academy of Sciences of the United States of America **100**(8): 4580-4585.

de Veer, M. J., M. Holko, M. Frevel, E. Walker, S. Der, J. M. Paranjape, R. H. Silverman and B. R. G. Williams (2001). "Functional classification of interferon-stimulated genes identified using microarrays." Journal of Leukocyte Biology **69**(6): 912-920.

DeKoter, R. P., J. C. Walsh, J. Anastasi, D. Friend, R. Stevens and H. Singh (1998). "PU.1 regulates multiple aspects of mast cell development." Blood **92**(10): 307A-307A.

DeKoter, R. P., J. C. Walsh and H. Singh (1998). "PU.1 regulates both cytokine-dependent proliferation and differentiation of granulocyte/macrophage progenitors." Embo Journal **17**(15): 4456-4468.

Dennis, G., B. T. Sherman, D. A. Hosack, J. Yang, W. Gao, H. C. Lane and R. A. Lempicki (2003). "DAVID: Database for annotation, visualization, and integrated discovery." Genome Biology **4**(9).

Deppmann, C. D., T. M. Thornton, F. E. Utama and E. J. Taparowsky (2003). "Phosphorylation of BATF regulates DNA binding: a novel mechanism for AP-1 (activator protein-1) regulation." Biochemical Journal **374**: 423-431.

Der, S. D., A. M. Zhou, B. R. G. Williams and R. H. Silverman (1998). "Identification of genes differentially regulated by interferon alpha, beta, or gamma using oligonucleotide arrays." Proceedings of the National Academy of Sciences of the United States of America **95**(26): 15623-15628.

Dersimonian, R. and N. Laird (1986). "Metaanalysis in Clinical-Trials." Controlled Clinical Trials **7**(3): 177-188.

Dey, A., L. Kim and W. Li (1999). "Gamma interferon induces expression of Mad1 gene in macrophage, which inhibits colony-stimulating factor-1-dependent mitogenesis." Journal of Cellular Biochemistry **72**(2): 232-241.

Dileepan, K. N., J. C. Page, Y. Li and D. J. Stechschulte (1995). "Direct activation of murine peritoneal macrophages for nitric oxide production and tumor cell killing by interferon-gamma." J Interferon Cytokine Res **15**(5): 387-394.

Dudoit, S., Y. H. Yang, M. J. Callow and T. P. Speed (2002). "Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments." Statistica Sinica **12**(1): 111-139.

Edgington, E. S. (1972). "Additive Method for Combining Probability Values from Independent Experiments." Journal of Psychology **80**(2): 351-&.

Efron, B., R. Tibshirani, J. D. Storey and V. Tusher (2001). "Empirical Bayes analysis of a microarray experiment." Journal of the American Statistical Association **96**(456): 1151-1160.

Egger, M. and G. D. Smith (1998). "Meta-analysis - Bias in location and selection of studies." British Medical Journal **316**(7124): 61-66.

Eide, E. J., M. F. Woolf, H. Kang, P. Woolf, W. Hurst, F. Camacho, E. L. Vielhaber, A. Giovanni and D. M. Virshup (2005). "Control of mammalian circadian rhythm by CKI epsilon-regulated proteasome-mediated PER2 degradation." Molecular and Cellular Biology **25**(7): 2795-2807.

Eisen, M. B., P. T. Spellman, P. O. Brown and D. Botstein (1998). "Cluster analysis and display of genome-wide expression patterns." Proceedings of the National Academy of Sciences of the United States of America **95**(25): 14863-14868.

Eisinga, R., R. Breitling and T. Heskes (2013). "The exact probability distribution of the rank product statistics for replicated experiments." FEBS Lett **587**(6): 677-682.

Farrar, M. A., J. AlberolaIla and R. M. Perlmutter (1996). "Activation of the Raf-1 kinase cascade by coumermycin-induced dimerization." Nature **383**(6596): 178-181.

Fasold, M., D. Langenberger, H. Binder, P. F. Stadler and S. Hoffmann (2011). "DARIO: a ncRNA detection and analysis tool for next-generation sequencing experiments." Nucleic Acids Research **39**: W112-W117.

Fielhaber, J. A., Y. S. Han, J. Tan, S. Xing, C. M. Biggs, K. B. Joung and A. S. Kristof (2009). "Inactivation of Mammalian Target of Rapamycin Increases STAT1 Nuclear Content and

Transcriptional Activity in alpha 4-and Protein Phosphatase 2A-dependent Fashion." Journal of Biological Chemistry **284**(36): 24341-24353.

Fisher, R. A. (1932). Statistical Methods For Research Workers. Edinburgh, Oliver & Boyd.

Flicek, P., I. Ahmed, M. R. Amode, D. Barrell, K. Beal, S. Brent, D. Carvalho-Silva, P. Clapham, G. Coates, S. Fairley, S. Fitzgerald, L. Gil, C. Garcia-Giron, L. Gordon, T. Hourlier, S. Hunt, T. Juettemann, A. K. Kahari, S. Keenan, M. Komorowska, E. Kulesha, I. Longden, T. Maurel, W. M. McLaren, M. Muffato, R. Nag, B. Overduin, M. Pignatelli, B. Pritchard, E. Pritchard, H. S. Riat, G. R. Ritchie, M. Ruffier, M. Schuster, D. Sheppard, D. Sobral, K. Taylor, A. Thormann, S. Trevanion, S. White, S. P. Wilder, B. L. Aken, E. Birney, F. Cunningham, I. Dunham, J. Harrow, J. Herrero, T. J. Hubbard, N. Johnson, R. Kinsella, A. Parker, G. Spudich, A. Yates, A. Zadissa and S. M. Searle (2013). "Ensembl 2013." Nucleic Acids Research **41**(Database issue): D48-55.

Fodor, S. P. A., J. L. Read, M. C. Pirrung, L. Stryer, A. T. Lu and D. Solas (1991). "Light-Directed, Spatially Addressable Parallel Chemical Synthesis." Science **251**(4995): 767-773.

Forrest, A. R. R. and P. Carninci (2009). "Whole genome transcriptome analysis." Rna Biology **6**(2): 107-112.

Forster, T., D. Roy and P. Ghazal (2003). "Experiments using microarray technology: limitations and standard operating procedures." Journal of Endocrinology **178**(2): 195-204.

Franceschini, A., D. Szklarczyk, S. Frankild, M. Kuhn, M. Simonovic, A. Roth, J. Lin, P. Minguez, P. Bork, C. von Mering and L. J. Jensen (2013). "STRING v9.1: protein-protein interaction networks, with increased coverage and integration." Nucleic Acids Research **41**(Database issue): D808-815.

Freeman, T. C., L. Goldovsky, M. Brosch, S. Van Dongen, P. Maziere, R. J. Grocock, S. Freilich, J. Thornton and A. J. Enright (2007). "Construction, visualisation, and clustering of transcription networks from Microarray expression data." Plos Computational Biology **3**(10): 2032-2042.

Fritzius, T. and K. Moelling (2008). "Akt- and foxo1-interacting WD-repeat-FYVE protein promotes adipogenesis." Embo Journal **27**(9): 1399-1410.

Gautier, L., L. Cope, B. M. Bolstad and R. A. Irizarry (2004). "affy - analysis of Affymetrix GeneChip data at the probe level." Bioinformatics **20**(3): 307-315.

Gavaghan, D. J., R. A. Moore and H. J. McQuay (2000). "An evaluation of homogeneity tests in meta-analyses in pain using simulations of individual patient data." Pain **85**(3): 415-424.

Gentleman, R. B. "geneplotter: Graphics related functions for Bioconductor."

Gentleman, R. C., V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. C. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Y. H. Yang and J. H. Zhang (2004). "Bioconductor: open software development for computational biology and bioinformatics." Genome Biology **5**(10).

Ghosh, D. B., T.R.; Rhodes, D; Chinnaiyan, A.M. (2003). "Statistical issues and methods for meta-analysis of microarray data: a case study in prostate cancer." Functional & Integrative Genomics **3**(4): 180-188.

Gillis, J. and P. Pavlidis (2012). ""Guilt by Association" Is the Exception Rather Than the Rule in Gene Networks." Plos Computational Biology **8**(3).

Glasbey, C. A., T. Forster and P. Ghazal (2007). "Estimation of expression levels in spotted microarrays with saturated pixels." Statistical Applications in Genetics and Molecular Biology **6**.

Glass, G. V. (1976). "Primary, secondary, and meta-analysis of research." Educational Researcher **5**: 3-8.

Golks, A., D. Brenner, P. H. Krammer and I. N. Lavrik (2006). "The c-FLIP-NH2 terminus (p22-FLIP) induces NF-kappa B activation." Journal of Experimental Medicine **203**(5): 1295-1305.

Gordon, S. (2003). "Alternative activation of macrophages." Nat Rev Immunol **3**(1): 23-35.

Greenhalgh, C. J. and D. J. Hilton (2001). "Negative regulation of cytokine signaling." Journal of Leukocyte Biology **70**(3): 348-356.

Grutzmann, R., H. Boriss, O. Ammerpohl, J. Luttges, H. Kalthoff, H. K. Schackert, G. Kloppel, H. D. Saeger and C. Pilarsky (2005). "Meta-analysis of microarray data on pancreatic cancer defines a set of commonly dysregulated genes." Oncogene **24**(32): 5079-5088.

Gyorffy, B. and R. Schafer (2009). "Meta-analysis of gene expression profiles related to relapse-free survival in 1,079 breast cancer patients." Breast Cancer Research and Treatment **118**(3): 433-441.

Hayashi, M., S. Shimba and M. Tezuka (2007). "Characterization of the molecular clock in mouse peritoneal macrophages." Biological & Pharmaceutical Bulletin **30**(4): 621-626.

He, G. A., Y. Ma, S. Y. Chou, H. H. Li, C. W. Yang, J. Z. Chuang, C. H. Sung and A. H. Ding (2011). "Role of CLIC4 in the host innate responses to bacterial lipopolysaccharide." European Journal of Immunology **41**(5): 1221-1230.

Hedges, L. V. O., I (1985). Statistical Methods for Meta-Analysis, Academic Press.

Heinzel, F. P., R. M. Rerko and A. M. Hujer (1998). "Underproduction of interleukin-12 in susceptible mice during progressive leishmaniasis is due to decreased CD40 activity." Cell Immunol **184**(2): 129-142.

Helming, L., J. Bose, J. Ehrchen, S. Schiebe, T. Frahm, R. Geffers, M. Probst-Kepper, R. Balling and A. Lengeling (2005). "1 alpha,25-dihydroxyvitamin D-3 is a potent suppressor of interferon gamma-mediated macrophage activation." Blood **106**(13): 4351-4358.

Hemmi, H., O. Takeuchi, S. Sato, M. Yamamoto, T. Kaisho, H. Sanjo, T. Kawai, K. Hoshino, K. Takeda and S. Akira (2004). "The roles of two I kappa B kinase-related kinases in lipopolysaccharide and double stranded RNA signaling and viral infection." Journal of Experimental Medicine **199**(12): 1641-1650.

Henderson, S. C., M. M. Kamdar and A. Bamezai (2002). "Ly-6A.2 expression regulates antigen-specific CD4(+) T cell proliferation and cytokine production." Journal of Immunology **168**(1): 118-126.

Hoffmann, R. (2008). "A wiki for the life sciences where authorship matters." Nature Genetics **40**(9): 1047-1051.

Hoffmann, R. and A. Valencia (2004). "A gene network for navigating the literature." Nature Genetics **36**(7): 664-664.

Hong, F. X. and R. Breitling (2008). "A comparison of meta-analysis methods for detecting differentially expressed genes in microarray experiments." Bioinformatics **24**(3): 374-382.

Hong, F. X., R. Breitling, C. W. McEntee, B. S. Wittner, J. L. Nemhauser and J. Chory (2006). "RankProd: a bioconductor package for detecting differentially expressed genes in meta-analysis." Bioinformatics **22**(22): 2825-2827.

Horvath, C. M. A., D (2013). Interferon gamma pathway, Sci. Signal.

Hou, W., X. Wang, L. Ye, L. Zhou, Z. Q. Yang, E. Riedel and W. Z. Ho (2009). "Lambda Interferon Inhibits Human Immunodeficiency Virus Type 1 Infection of Macrophages." Journal of Virology **83**(8): 3834-3842.

Hsu, J. C. (1996). Multiple Comparisons: Theory and methods. London, Chapman & Hall.

Hu, S. H., Z. Xie, A. Onishi, X. P. Yu, L. Z. Jiang, J. Lin, H. S. Rho, C. Woodard, H. Wang, J. S. Jeong, S. Y. Long, X. F. He, H. Wade, S. Blackshaw, J. Qian and H. Zhu (2009). "Profiling the Human Protein-DNA Interactome Reveals ERK2 as a Transcriptional Repressor of Interferon Signaling." Cell **139**(3): 610-622.

Hu, X., J. Chen, L. Wang and L. B. Ivashkiv (2007). "Crosstalk among Jak-STAT, Toll-like receptor, and ITAM-dependent pathways in macrophage activation." J Leukoc Biol **82**(2): 237-243.

Hu, X. Y., P. K. Paik, J. Chen, A. Yarilina, L. Kockeritz, T. T. Lu, J. R. Woodgett and L. B. Ivashkiv (2006). "IFN-gamma suppresses IL-10 production and synergizes with TLR2 by regulating GSK3 and CREB/AP-1 proteins." Immunity **24**(5): 563-574.

Huang, T., K. Tu, Y. Shyr, C. C. Wei, L. Xie and Y. X. Li (2008). "The prediction of interferon treatment effects based on time series microarray gene expression profiles." Journal of Translational Medicine **6**.

Huang, Z., X. Zhou, A. C. Nicholson, A. M. Gotto, D. P. Hajjar and J. Han (2008). "Activation of peroxisome proliferator-activated receptor-alpha in mice induces expression of the hepatic low-density lipoprotein receptor." British Journal of Pharmacology **155**(4): 596-605.

Huber, W., A. von Heydebreck, H. Sultmann, A. Poustka and M. Vingron (2002). "Variance stabilization applied to microarray data calibration and to the quantification of differential expression." Bioinformatics **18 Suppl 1**: S96-104.

Hughes, A. L., D. W. Powell, M. Bard, J. Eckstein, R. Barbuch, A. J. Link and P. J. Espenshade (2007). "Dap1/PGRMC1 binds and regulates cytochrome P450 enzymes." Cell Metabolism **5**(2): 143-149.

Husain, Z., N. Holodick, C. Day, I. Szymanski and C. A. Alper (2006). "Increased apoptosis of CD20(+) IgA(+) B cells is the basis for IgA deficiency: The molecular mechanism for correction in vitro by IL-10 and CD40L." Journal of Clinical Immunology **26**(2): 113-125.

Ikeda, F., C. M. Hecker, A. Rozenknop, R. D. Nordmeier, V. Rogov, K. Hofmann, S. Akira, V. Dotsch and I. Dikic (2007). "Involvement of the ubiquitin-like domain of TBK1/IKK-i kinases in regulation of IFN-inducible genes." Embo Journal **26**(14): 3451-3462.

Irizarry, R. A., B. M. Bolstad, F. Collin, L. M. Cope, B. Hobbs and T. P. Speed (2003). "Summaries of Affymetrix GeneChip probe level data." Nucleic Acids Research **31**(4): e15.

Irizarry, R. A., B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf and T. P. Speed (2003). "Exploration, normalization, and summaries of high density oligonucleotide array probe level data." Biostatistics **4**(2): 249-264.

Irizarry, R. A., D. Warren, F. Spencer, I. F. Kim, S. Biswal, B. C. Frank, E. Gabrielson, J. G. N. Garcia, J. Geoghegan, G. Germino, C. Griffin, S. C. Hilmer, E. Hoffman, A. E. Jedlicka, E. Kawasaki, F. Martinez-Murillo, L. Morsberger, H. Lee, D. Petersen, J. Quackenbush, A. Scott, M. Wilson, Y. Q. Yang, S. Q. Ye and W. Yu (2005). "Multiple-laboratory comparison of microarray platforms." Nature Methods **2**(5): 345-349.

Isaacs, A. and J. Lindenmann (1957). "Virus interference. I. The interferon." Proc R Soc Lond B Biol Sci **147**(927): 258-267.

Jaffrezic, A. R. G. M. F. (2013). Differential meta-analysis of RNA-seq data from multiple studies, arXiv.org.

Jeanmougin, M., A. de Reynies, L. Marisa, C. Paccard, G. Nuel and M. Guedj (2010). "Should We Abandon the t-Test in the Analysis of Gene Expression Microarray Data: A Comparison of Variance Modeling Strategies." Plos One **5**(9).

Jesenberger, V., K. J. Procyk, J. Ruth, M. Schreiber, H. C. Theussl, E. F. Wagner and M. Baccarini (2001). "Protective role of Raf-1 in Salmonella-induced macrophage apoptosis." Journal of Experimental Medicine **193**(3): 353-364.

Jornsten, R., H. Y. Wang, W. J. Welsh and M. Ouyang (2005). "DNA microarray data imputation and significance analysis of differential expression." Bioinformatics **21**(22): 4155-4161.

Kalai, M., M. Lamkanfi, G. Denecker, M. Boogmans, S. Lippens, A. Meeus, W. Declercq and P. Vandenabeele (2003). "Regulation of the expression and processing of caspase-12." Journal of Cell Biology **162**(3): 457-467.

Karki, S., M. M. Li, J. W. Schoggins, S. Tian, C. M. Rice and M. R. MacDonald (2012). "Multiple interferon stimulated genes synergize with the zinc finger antiviral protein to mediate anti-alphavirus activity." Plos One **7**(5): e37398.

Kashatus, D., P. Cogswell and A. S. Baldwin (2006). "Expression of the Bcl-3 proto-oncogene suppresses p53 activation." Genes & Development **20**(2): 225-235.

Kasprzyk, A. (2011). "BioMart: driving a paradigm change in biological data management." Database-the Journal of Biological Databases and Curation.

Kauffmann, A., R. Gentleman and W. Huber (2009). "arrayQualityMetrics--a bioconductor package for quality assessment of microarray data." Bioinformatics **25**(3): 415-416.

Keegan, K. P., S. Pradhan, J. P. Wang and R. Allada (2007). "Meta-analysis of Drosophila circadian microarray studies identifies a novel set of rhythmically expressed genes." Plos Computational Biology **3**(11): 2087-2110.

Kerr, M. K. and G. A. Churchill (2001). "Statistical design and the analysis of gene expression microarray data." Genet Res **77**(2): 123-128.

Kerr, M. K., M. Martin and G. A. Churchill (2000). "Analysis of variance for gene expression microarray data." J Comput Biol **7**(6): 819-837.

Khan, J., L. H. Saal, M. L. Bittner, Y. D. Chen, J. M. Trent and P. S. Meltzer (1999). "Expression profiling in cancer using cDNA microarrays." Electrophoresis **20**(2): 223-229.

Kim, H., G. H. Golub and H. Park (2005). "Missing value estimation for DNA microarray gene expression data: local least squares imputation." Bioinformatics **21**(2): 187-198.

Kishida, T., I. Kostetskii, Z. B. Zhang, F. Martinez, P. Liu, S. U. Walkley, N. K. Dwyer, E. J. Blanchette-Mackie, G. L. Radice and J. F. Strauss (2004). "Targeted mutation of the MLN64 START domain causes only modest alterations in cellular sterol metabolism." Journal of Biological Chemistry **279**(18): 19276-19285.

Kropp, K. A., K. A. Robertson, G. Sing, S. Rodriguez-Martin, M. Blanc, P. Lacaze, M. F. B. N. Hassim, M. R. Khondoker, A. Busche, P. Dickinson, T. Forster, B. Strobl, M. Mueller, S. Jonjic, A. Angulo and P. Ghazal (2011). "Reversible Inhibition of Murine Cytomegalovirus Replication by Gamma Interferon (IFN-gamma) in Primary Macrophages Involves a Primed Type I IFN-Signaling Subnetwork for Full Establishment of an Immediate-Early Antiviral State." Journal of Virology **85**(19): 10286-10299.

Kulesh, D. A., D. R. Clive, D. S. Zarlenga and J. J. Greene (1987). "Identification of Interferon-Modulated Proliferation-Related Cdna Sequences." Proceedings of the National Academy of Sciences of the United States of America **84**(23): 8453-8457.

Kupershmidt, I., Q. J. Su, A. Grewal, S. Sundaresh, I. Halperin, J. Flynn, M. Shekar, H. Wang, J. Park, W. W. Cui, G. D. Wall, R. Wisotzkey, S. Alag, S. Akhtari and M. Ronaghi (2010). "Ontology-Based Meta-Analysis of Global Collections of High-Throughput Public Data." Plos One **5**(9).

Lashkari, D. A., J. H. McCusker and R. W. Davis (1997). "Whole genome analysis: Experimental access to all genome sequenced segments through larger-scale efficient oligonucleotide synthesis and PCR." Proceedings of the National Academy of Sciences of the United States of America **94**(17): 8945-8947.

Leakey, A. K., G. C. Ulett and R. G. Hirst (1998). "BALB/c and C57Bl/6 mice infected with virulent Burkholderia pseudomallei provide contrasting animal models for the acute and chronic forms of human melioidosis." Microb Pathog **24**(5): 269-275.

Lei, C. Q., Y. Zhang, T. Xia, L. Q. Jiang, B. Zhong and H. B. Shu (2013). "FoxO1 negatively regulates cellular antiviral response by promoting degradation of IRF3." Journal of Biological Chemistry.

Lewis, S. and M. Clarke (2001). "Forest plots: trying to see the wood and the trees." British Medical Journal **322**(7300): 1479-1480.

Li, J. and R. Tibshirani (2013). "Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-Seq data." Stat Methods Med Res **22**(5): 519-536.

Li, L. S., C. Yee and J. A. Beavo (1999). "CD3- and CD28-dependent induction of PDE7 required for T cell activation." Science **283**(5403): 848-851.

Liekens, A. M. L., J. De Knijf, W. Daelemans, B. Goethals, P. De Rijk and J. Del-Favero (2011). "BioGraph: unsupervised biomedical knowledge discovery via automated hypothesis generation." Genome Biology **12**(6).

Lu, L. F., T. H. Thai, D. P. Calado, A. Chaudhry, M. Kubo, K. Tanaka, G. B. Loeb, H. Lee, A. Yoshimura, K. Rajewsky and A. Y. Rudensky (2009). "Foxp3-Dependent MicroRNA155

Confers Competitive Fitness to Regulatory T Cells by Targeting SOCS1 Protein." Immunity **30**(1): 80-91.

Lusa, L. G., R; Ruschhaupt, M "GeneMeta: MetaAnalysis for High Throughput Experiments."

Mack, K. D., M. Von Goetz, M. Lin, M. Venegas, J. Barnhart, Y. Lu, B. Lamar, R. Stull, C. Silvin, P. Owings, F. Y. Bih and A. Abo (2005). "Functional identification of kinases essential for T-cell activation through a genetic suppression screen." Immunology Letters **96**(1): 129-145.

Mackaness, G. B. (1964). "Immunological Basis of Acquired Cellular Resistance." Journal of Experimental Medicine **120**(1): 105-&.

Mackaness, G. B. (1969). "Influence of Immunologically Committed Lymphoid Cells on Macrophage Activity in Vivo." Journal of Experimental Medicine **129**(5): 973-&.

Maglott, D., J. Ostell, K. D. Pruitt and T. Tatusova (2011). "Entrez Gene: gene-centered information at NCBI." Nucleic Acids Research **39**: D52-D57.

Malik, M., K. Jividen, V. C. Padmakumar, C. Cataisson, L. W. Li, J. Lee, O. M. Z. Howard and S. H. Yuspa (2012). "Inducible NOS-induced chloride intracellular channel 4 (CLIC4) nuclear translocation regulates macrophage deactivation." Proceedings of the National Academy of Sciences of the United States of America **109**(16): 6130-6135.

Mardis, E. R. (2008). "Next-generation DNA sequencing methods." Annu Rev Genomics Hum Genet **9**: 387-402.

Marot, G. and C. D. Mayer (2009). "Sequential Analysis for Microarray Data Based on Sensitivity and Meta-Analysis." Statistical Applications in Genetics and Molecular Biology **8**(1).

Martin, G. R. W. I. R. s. c. a. o. d. c. b. B. B. a. L. B. a. R. G. a. W. H. A. L. a. T. L. a. (2011). "gplots: Various R programming tools for plotting data."

Martinez, F. O., S. Gordon, M. Locati and A. Mantovani (2006). "Transcriptional profiling of the human monocyte-to-macrophage differentiation and polarization: new molecules and patterns of gene expression." Journal of Immunology **177**(10): 7303-7311.

Mashima, R., K. Saeki, D. Aki, Y. Minoda, H. Takaki, T. Sanada, T. Kobayashi, H. Aburatani, Y. Yamanashi and A. Yoshimura (2005). "FLN29, a novel interferon- and LPS-inducible gene acting as a negative regulator of toll-like receptor signaling." Journal of Biological Chemistry **280**(50): 41289-41297.

Maskos, U. and E. M. Southern (1992). "Oligonucleotide Hybridizations on Glass Supports - a Novel Linker for Oligonucleotide Synthesis and Hybridization Properties of Oligonucleotides Synthesized Insitu." Nucleic Acids Research **20**(7): 1679-1684.

Maskos, U. and E. M. Southern (1993). "A Study of Oligonucleotide Reassociation Using Large Arrays of Oligonucleotides Synthesized on a Glass Support." Nucleic Acids Research **21**(20): 4663-4669.

Melchjorsen, J., J. Siren, K. Julkunen, S. R. Paludan and S. Matikainen (2006). "Induction of cytokine expression by herpes simplex virus in human monocyte-derived macrophages and dendritic cells is dependent on virus replication and is counteracted by ICP27 targeting NF-kappa B and IRF-3." Journal of General Virology **87**: 1099-1108.

Metchnikoff, E. (1905). Immunity in Infective Diseases, Cambridge University Press.

Metzker, M. L. (2010). "Applications of Next-Generation Sequencing Sequencing Technologies - the Next Generation." Nature Reviews Genetics **11**(1): 31-46.

Mills, C. D., K. Kincaid, J. M. Alt, M. J. Heilman and A. M. Hill (2000). "M-1/M-2 macrophages and the Th1/Th2 paradigm." J Immunol **164**(12): 6166-6173.

Mooney, M., J. Bond, N. Monks, E. Eugster, D. Cherba, P. Berlinski, S. Kamerling, K. Marotti, H. Simpson, T. Rusk, W. Tembe, C. Legendre, H. Benson, W. Liang and C. P. Webb (2013). "Comparative RNA-Seq and microarray analysis of gene expression changes in B-cell lymphomas of Canis familiaris." PLoS One **8**(4): e61088.

Mortazavi, A., B. A. Williams, K. McCue, L. Schaeffer and B. Wold (2008). "Mapping and quantifying mammalian transcriptomes by RNA-Seq." Nat Methods **5**(7): 621-628.

Muller, U., U. Steinhoff, L. F. L. Reis, S. Hemmi, J. Pavlovic, R. M. Zinkernagel and M. Aguet (1994). "Functional-Role of Type-I and Type-Ii Interferons in Antiviral Defense." Science **264**(5167): 1918-1921.

Naslavsky, N., J. Rahajeng, D. Rapaport, M. Horowitz and S. Caplan (2007). "EHD1 regulates cholesterol homeostasis and lipid droplet storage." Biochemical and Biophysical Research Communications **357**(3): 792-799.

Naslavsky, N., J. Rahajeng, M. Sharma, M. Jovic and S. Caplan (2006). "Interactions between EHD proteins and Rab11-FIP2: A role for EHD3 in early endosomal transport." Molecular Biology of the Cell **17**(1): 163-177.

Nathan, C. (1991). "Mechanisms and Modulation of Macrophage Activation." Behring Institute Mitteilungen, No 88 **88**: 200-207.

Nathan, C. F., H. W. Murray, M. E. Wiebe and B. Y. Rubin (1983). "Identification of Interferon-Gamma as the Lymphokine That Activates Human Macrophage Oxidative-Metabolism and Anti-Microbial Activity." Journal of Experimental Medicine **158**(3): 670-689.

Neess, D., P. Kiilerich, M. B. Sandberg, T. Helledie, R. Nielsen and S. Mandrup (2006). "ACBP - a PPAR and SREBP modulated housekeeping gene." Molecular and Cellular Biochemistry **284**(1-2): 149-157.

Ng, S. L., B. A. Friedman, S. Schmid, J. Gertz, R. M. Myers, B. R. tenOever and T. Maniatis (2011). "I kappa B kinase epsilon (IKK epsilon) regulates the balance between type I and type II interferon responses." Proceedings of the National Academy of Sciences of the United States of America **108**(52): 21170-21175.

Nikaido, K. T. I. (2013). metaSeq: Meta-analysis of RNA-Seq count data in multiple studies. R package version 1.2.2.

Nolte, I., D. Jeckel, F. T. Wieland and K. Sohn (2000). "Localization and topology of ratp28, a member of a novel family of putative steroid-binding proteins." Biochimica Et Biophysica Acta-Protein Structure and Molecular Enzymology **1543**(1): 123-130.

Oba, S., M. Sato, I. Takemasa, M. Monden, K. Matsubara and S. Ishii (2003). "A Bayesian missing value estimation method for gene expression profile data." Bioinformatics **19**(16): 2088-2096.

Oh, Y. M., J. K. Kim, Y. Choi, S. Choi and J. Y. Yoo (2009). "Prediction and Experimental Validation of Novel STAT3 Target Genes in Human Cancer Cells." Plos One **4**(9).

Oldenborg, P. A., H. D. Gresham and F. P. Lindberg (2001). "CD47-signal regulatory protein alpha (SIRP alpha) regulates Fc gamma and complement receptor-mediated phagocytosis." Journal of Experimental Medicine **193**(7): 855-861.

Oswald, I. P., S. Afroun, D. Bray, J. F. Petit and G. Lemaire (1992). "Low response of BALB/c macrophages to priming and activating signals." J Leukoc Biol **52**(3): 315-322.

Ouchi, T., S. W. Lee, M. Ouchi, S. A. Aaronson and C. M. Horvath (2000). "Collaboration of signal transducer and activator of transcription 1 (STAT1) and BRCA1 in differential regulation of IFN-gamma target genes." Proceedings of the National Academy of Sciences of the United States of America **97**(10): 5208-5213.

Ozbudak, E. M., M. Thattai, I. Kurtser, A. D. Grossman and A. van Oudenaarden (2002). "Regulation of noise in the expression of a single gene." Nature Genetics **31**(1): 69-73.

Pagan, J. K., F. G. Wylie, S. Joseph, C. Widberg, N. J. Bryant, D. E. James and J. L. Stow (2003). "The t-SNARE syntaxin 4 is regulated during macrophage activation to function in membrane traffic and cytokine secretion." Current Biology **13**(2): 156-160.

Pandey, A. K., N. Munjal and M. Datta (2010). "Gene Expression Profiling and Network Analysis Reveals Lipid and Steroid Metabolism to Be the Most Favored by TNF alpha in HepG2 Cells." Plos One **5**(2).

Parmigiani, G., E. S. Garrett-Mayer, R. Anbazhagan and E. Gabrielson (2004). "A cross-study comparison of gene expression studies for the molecular classification of lung cancer." Clinical Cancer Research **10**(9): 2922-2927.

Perry, A. K., E. K. Chow, J. B. Goodnough, W. C. Yeh and G. Cheng (2004). "Differential requirement for TANK-binding kinase-1 in type I interferon responses to toll-like receptor activation and viral infection." Journal of Experimental Medicine **199**(12): 1651-1658.

Platanias, L. C. (2005). Mechanisms of type-I- and type-II-interferon-mediated signalling. Nature Reviews Immunology**:** 375-386.

Popkin, D. L., M. A. Watson, E. Karaskov, G. P. Dunn, R. Bremner and H. W. Virgin (2003). "Murine cytomegalovirus paralyzes macrophages by blocking IFN gamma-induced promoter assembly." Proceedings of the National Academy of Sciences of the United States of America **100**(24): 14309-14314.

Quackenbush, J. (2001). "Computational analysis of microarray data." Nature Reviews Genetics **2**(6): 418-427.

Quackenbush, J. (2002). "Microarray data normalization and transformation." Nature Genetics **32**: 496-501.

Quackenbush, J. (2003). "Genomics. Microarrays--guilt by association." Science **302**(5643): 240-241.

R Development Core Team (2009). R: A language and environment for statistical computing, R Foundation for Statistical Computing.

Ramana, C. V., M. P. Gil, R. D. Schreiber and G. R. Stark (2002). "Stat1-dependent and -independent pathways in IFN-gamma-dependent signaling." Trends in Immunology **23**(2): 96-101.

Ramasamy, A., A. Mondry, C. C. Holmes and D. G. Altman (2008). "Key Issues in Conducting a Meta-Analysis of Gene Expression Microarray Datasets." Plos Medicine **5**(9): 1320-1332.

Rane, S. G. and E. P. Reddy (2000). "Janus kinases: components of multiple signaling pathways." Oncogene **19**(49): 5662-5679.

Rathinam, V. A. K. and K. A. Fitzgerald (2011). "Innate immune sensing of DNA viruses." Virology **411**(2): 153-162.

Raza, S., N. McDerment, P. A. Lacaze, K. Robertson, S. Watterson, Y. Chen, M. Chisholm, G. Eleftheriadis, S. Monk, M. O'Sullivan, A. Turnbull, D. Roy, A. Theocharidis, P. Ghazal and T. C. Freeman (2010). "Construction of a large scale integrated map of macrophage pathogen recognition and effector systems." Bmc Systems Biology **4**.

Raza, S., K. A. Robertson, P. A. Lacaze, D. Page, A. J. Enright, P. Ghazal and T. C. Freeman (2008). "A logic-based diagram of signalling pathways central to macrophage activation." Bmc Systems Biology **2**.

Rederstorff, M., S. H. Bernhart, A. Tanzer, M. Zywicki, K. Perfler, M. Lukasser, I. L. Hofacker and A. Huttenhofer (2010). "RNPomics: defining the ncRNA transcriptome by cDNA library generation from ribonucleo-protein particles." Nucleic Acids Res **38**(10): e113.

Reiner, A., D. Yekutieli and Y. Benjamini (2003). "Identifying differentially expressed genes using false discovery rate controlling procedures." Bioinformatics **19**(3): 368-375.

Reverter, A. and E. K. Chan (2008). "Combining partial correlation and an information theory approach to the reversed engineering of gene co-expression networks." Bioinformatics **24**(21): 2491-2497.

Rhodes, D. R., T. R. Barrette, M. A. Rubin, D. Ghosh and A. M. Chinnaiyan (2002). "Meta-analysis of microarrays: Interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer." Cancer Research **62**(15): 4427-4433.

Rhodes, D. R., J. J. Yu, K. Shanker, N. Deshpande, R. Varambally, D. Ghosh, T. Barrette, A. Pandey and A. M. Chinnaiyan (2004). "Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression." Proceedings of the National Academy of Sciences of the United States of America **101**(25): 9309-9314.

Richards, K. H. and A. Macdonald (2011). "Putting the brakes on the anti-viral response: negative regulators of type I interferon (IFN) production." Microbes and Infection **13**(4): 291-302.

Riemann, M., R. Endres, S. Liptay, K. Pfeffer and R. M. Schmid (2005). "The I kappa B protein Bcl-3 negatively regulates transcription of the IL-10 gene in macrophages." Journal of Immunology **175**(6): 3560-3568.

Robinson, M. D., D. J. McCarthy and G. K. Smyth (2010). "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data." Bioinformatics **26**(1): 139-140.

Rouse, R. and G. Hardiman (2003). "Microarray technology - an intellectual property retrospective." Pharmacogenomics **4**(5): 623-632.

Rubin, D. B. (1976). "Inference and Missing Data." Biometrika **63**(3): 581-590.

Safran, M., I. Dalah, J. Alexander, N. Rosen, T. I. Stein, M. Shmoish, N. Nativ, I. Bahir, T. Doniger, H. Krug, A. Sirota-Madi, T. Olender, Y. Golan, G. Stelzer, A. Harel and D. Lancet (2010). "GeneCards Version 3: the human gene integrator." Database-the Journal of Biological Databases and Curation.

Samarajiwa, S. A., S. Forster, K. Auchettl and P. J. Hertzog (2009). "INTERFEROME: the database of interferon regulated genes." Nucleic Acids Research **37**: D852-D857.

Samuel, C. E. (2001). "Antiviral actions of interferons." Clin Microbiol Rev **14**(4): 778-809, table of contents.

Sanada, T., G. Takaesu, R. Mashima, R. Yoshida, T. Kobayashi and A. Yoshimura (2008). "FLN29 deficiency reveals its negative regulatory role in the Toll-like receptor (TLR) and retinoic acid-inducible gene I (RIG-I)-like helicase signaling pathway." Journal of Biological Chemistry **283**(49): 33858-33864.

Santos, J. L., A. A. Andrade, A. A. Dias, C. A. Bonjardim, L. F. Reis, S. M. Teixeira and M. F. Horta (2006). "Differential sensitivity of C57BL/6 (M-1) and BALB/c (M-2) macrophages to the stimuli of IFN-gamma/LPS for the production of NO: correlation with iNOS mRNA and protein expression." J Interferon Cytokine Res **26**(9): 682-688.

Sather, S., K. D. Kenyon, J. B. Lefkowitz, X. Y. Liang, B. C. Varnum, P. M. Henson and D. K. Graham (2007). "A soluble form of the Mer receptor tyrosine kinase inhibits macrophage clearance of apoptotic cells and platelet aggregation." Blood **109**(3): 1026-1033.

Scharpf, R. B., H. Tjelmeland, G. Parmigiani and A. B. Nobel (2009). "A Bayesian Model for Cross-Study Differential Gene Expression." Journal of the American Statistical Association **104**(488): 1295-1310.

Scheel, I., M. Aldrin, I. K. Glad, R. Sorum, H. Lyng and A. Frigessi (2005). "The influence of missing value imputation on detection of differentially expressed genes from microarray data." Bioinformatics **21**(23): 4272-4279.

Schena, M., D. Shalon, R. W. Davis and P. O. Brown (1995). "Quantitative Monitoring of Gene-Expression Patterns with a Complementary-DNA Microarray." Science **270**(5235): 467-470.

Schomacker, H., R. M. Hebner, J. Boonyaratanakornkit, S. Surman, E. Amaro-Carambot, P. L. Collins and A. C. Schmidt (2012). "The C Proteins of Human Parainfluenza Virus Type 1 Block IFN Signaling by Binding and Retaining Stat1 in Perinuclear Aggregates at the Late Endosome." Plos One **7**(2).

Schroder, K., P. J. Hertzog, T. Ravasi and D. A. Hume (2004). Interferon-gamma: an overview of signals, mechanisms and functions. J Leukoc Biol**:** 163-189.

Schwarz, E. M., P. Krimpenfort, A. Berns and I. M. Verma (1997). "Immunological defects in mice with a targeted disruption in Bcl-3." Genes & Development **11**(2): 187-197.

Sean, D. and P. S. Meltzer (2007). "GEOquery: a bridge between the gene expression omnibus (GEO) and BioConductor." Bioinformatics **23**(14): 1846-1847.

Senga, T., T. Iwamoto, S. E. Humphrey, T. Yokota, E. J. Taparowsky and M. Hamaguchi (2002). "Stat3-dependent induction of BATF in M1 mouse myeloid leukemia cells." Oncogene **21**(53): 8186-8191.

Sica, A. and A. Mantovani (2012). "Macrophage plasticity and polarization: in vivo veritas." Journal of Clinical Investigation **122**(3): 787-795.

Sidak, Z. (1967). "Rectangular Confidence Regions for Means of Multivariate Normal Distributions." Journal of the American Statistical Association **62**(318): 626-&.

Simes, R. J. (1986). "An Improved Bonferroni Procedure for Multiple Tests of Significance." Biometrika **73**(3): 751-754.

Simpson, R. J. S. and K. Pearson (1904). "Report on certain enteric fever inoculation statistics." British Medical Journal **1904**: 1243-1246.

Sims, A. H., G. J. Smethurst, Y. Hey, M. J. Okoniewski, S. D. Pepper, A. Howell, C. J. Miller and R. B. Clarke (2008). "The removal of multiplicative, systematic bias allows integration of breast cancer gene expression datasets - improving meta-analysis and prediction of prognosis." Bmc Medical Genomics **1**.

Sjin, R. M. T. T., K. Krishnaraju, B. Hoffman and D. A. Liebermann (2002). "Transcriptional regulation of myeloid differentiation primary response (MyD) genes during myeloid differentiation is mediated by nuclear factor Y." Blood **100**(1): 80-88.

Smyth, G. K. (2004). "Linear models and empirical bayes methods for assessing differential expression in microarray experiments." Stat Appl Genet Mol Biol **3**: Article3.

Smyth, G. K. (2005). Limma: Linear models for microarray data. Statistics for Biology and Health. R. C. Gentleman, VJ; Huber, W; Irizarry, RA; Dudoit, S. New York, Springer**:** 397-420.

Soneson, C. and M. Delorenzi (2013). "A comparison of methods for differential expression analysis of RNA-seq data." BMC Bioinformatics **14**: 91.

Southern, E. M. (1975). "Detection of Specific Sequences among DNA Fragments Separated by Gel-Electrophoresis." Journal of Molecular Biology **98**(3): 503-&.

Southern, E. M., S. C. Casegreen, J. K. Elder, M. Johnson, K. U. Mir, L. Wang and J. C. Williams (1994). "Arrays of Complementary Oligonucleotides for Analyzing the Hybridization Behavior of Nucleic-Acids." Nucleic Acids Research **22**(8): 1368-1373.

Stacklies, W., H. Redestig, M. Scholz, D. Walther and J. Selbig (2007). "pcaMethods - a bioconductor package providing PCA methods for incomplete data." Bioinformatics **23**(9): 1164-1167.

Stein, M., S. Keshav, N. Harris and S. Gordon (1992). "Interleukin 4 potently enhances murine macrophage mannose receptor activity: a marker of alternative immunologic macrophage activation." Journal of Experimental Medicine **176**(1): 287-292.

Storey, J. D. and R. Tibshirani (2003). "Statistical significance for genomewide studies." Proceedings of the National Academy of Sciences of the United States of America **100**(16): 9440-9445.

Stouffer, S. A. (1949). The American Soldier, Princeton University Press.

Strauss, J. F., P. Liu, L. K. Christenson and H. Watari (2002). "Sterols and intracellular vesicular trafficking: lessons from the study of NPC1." Steroids **67**(12): 947-951.

Suchanek, M., A. Radzikowska and C. Thiele (2005). "Photo-leucine and photo-methionine allow identification of protein-protein interactions in living cells." Nature Methods **2**(4): 261-267.

Sugiyama, T., H. Kumagai, Y. Morikawa, Y. Wada, A. Sugiyama, K. Yasuda, N. Yokoi, S. Tamura, T. Kojima, T. Nosaka, E. Senba, S. Kimura, T. Kadowaki, T. Kodama and T. Kitamura (2000). "A novel low-density lipoprotein receptor-related protein mediating cellular uptake of apolipoprotein E-enriched beta-VLDL in vitro." Biochemistry **39**(51): 15817-15825.

Sui, S. J. H., D. L. Fulton, D. J. Arenillas, A. T. Kwon and W. W. Wasserman (2007). "oPOSSUM: integrated tools for analysis of regulatory motif over-representation." Nucleic Acids Research **35**: W245-W252.

Suriano, A. R., A. N. Sanford, N. Kim, M. Oh, S. Kennedy, M. J. Henderson, K. Dietzmann and K. E. Sullivan (2005). "GCF2/LRRFIP1 represses tumor necrosis factor alpha expression." Molecular and Cellular Biology **25**(20): 9073-9081.

Takeuchi, O., H. Hemmi and S. Akira (2004). "Interferon response induced by Toll-like receptor signaling." Journal of Endotoxin Research **10**(4): 252-256.

Taniguchi, T. and A. Takaoka (2002). "The interferon-alpha/beta system in antiviral responses: a multimodal machinery of gene regulation by the IRF family of transcription factors." Current Opinion in Immunology **14**(1): 111-116.

Tanuma, N., K. Nakamura, H. Shima and K. Kikuchi (2000). "Protein-tyrosine phosphatase PTP epsilon C inhibits Jak-STAT signaling and differentiation induced by interleukin-6 and leukemia inhibitory factor in M1 leukemia cells." Journal of Biological Chemistry **275**(36): 28216-28221.

Tanuma, N., H. Shima, K. Nakamura and K. Kikuchi (2001). "Protein tyrosine phosphatase epsilon C selectively inhibits interleukin-6-and interleukin-10-induced JAK-STAT signaling." Blood **98**(10): 3030-3034.

Thomas, K. E., C. L. Galligan, R. D. Newman, E. N. Fish and S. N. Vogel (2006). "Contribution of interferon-beta to the murine macrophage response to the toll-like receptor 4 agonist, lipopolysaccharide." Journal of Biological Chemistry **281**(41): 31119-31130.

Thomassen, M., Q. H. Tan and T. A. Kruse (2009). "Gene expression meta-analysis identifies chromosomal regions and candidate genes involved in breast cancer metastasis." Breast Cancer Research and Treatment **113**(2): 239-249.

Thompson, M. R., J. J. Kaminski, E. A. Kurt-Jones and K. A. Fitzgerald (2011). "Pattern Recognition Receptors and the Innate Immune Response to Viral Infection." Viruses-Basel **3**(6): 920-940.

Thornton, T. M., A. J. Zullo, K. L. Williams and E. J. Taparowsky (2006). "Direct manipulation of activator protein-1 controls thymocyte proliferation in vitro." European Journal of Immunology **36**(1): 160-169.

Tripathy, S. P. and G. B. Mackaness (1969). "Effect of Cytotoxic Agents on Passive Transfer of Cell-Mediated Immunity." Journal of Experimental Medicine **130**(1): 17-&.

Troyanskaya, O., M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein and R. B. Altman (2001). "Missing value estimation methods for DNA microarrays." Bioinformatics **17**(6): 520-525.

Tseng, G. C., D. Ghosh and E. Feingold (2012). "Comprehensive literature review and statistical considerations for microarray meta-analysis." Nucleic Acids Research **40**(9): 3785-3799.

Tussiwand, R., W. L. Lee, T. L. Murphy, M. Mashayekhi, K. C. Wumesh, J. C. Albring, A. T. Satpathy, J. A. Rotondo, B. T. Edelson, N. M. Kretzer, X. D. Wu, L. A. Weiss, E. Glasmacher, P. Li, W. Liao, M. Behnke, S. S. K. Lam, C. T. Aurthur, W. J. Leonard, H. Singh, C. L. Stallings, L. D. Sibley, R. D. Schreiber and K. M. Murphy (2012). "Compensatory dendritic cell development mediated by BATF-IRF interactions." Nature **490**(7421): 502-+.

Vadiveloo, P. K., G. Vairo, P. Hertzog, I. Kola and J. A. Hamilton (2000). "Role of type I interferons during macrophage activation by lipopolysaccharide." Cytokine **12**(11): 1639-1646.

van Erp, K., K. Dach, I. Koch, R. Heesemann and R. Hoffmann (2006). "Role of strain differences on host resistance and the transcriptional response of macrophages to infection with Yersinia enterocolitica." Physiological Genomics **25**(1): 75-84.

Vanfurth, R., W. G. Spector, Z. A. Cohn, J. G. Hirsch, Langevoo.Hl and J. H. Humphrey (1972). "Mononuclear Phagocyte System - New Classification of Macrophages, Monocytes, and Their Precursor Cells." Bulletin of the World Health Organization **46**(6): 845-&.

Velez, D. R., W. F. Hulme, J. L. Myers, J. B. Weinberg, M. C. Levesque, M. E. Stryjewski, E. Abbate, R. Estevan, S. G. Patillo, J. R. Gilbert, C. D. Hamilton and W. K. Scott (2009). "NOS2A, TLR4, and IFNGR1 interactions influence pulmonary tuberculosis susceptibility in African-Americans." Human Genetics **126**(5): 643-653.

Vert, G., J. L. Nemhauser, N. Geldner, F. X. Hong and J. Chory (2005). "Molecular mechanisms of steroid hormone signaling in plants." Annual Review of Cell and Developmental Biology **21**: 177-201.

Vilcek, J. (2003). Novel interferons. Nat Immunol**:** 8-9.

Waddell, S. J., S. J. Popper, K. H. Rubins, M. J. Griffiths, P. O. Brown, M. Levin and D. A. Relman (2010). "Dissecting Interferon-Induced Transcriptional Programs in Human Peripheral Blood Cells." Plos One **5**(3).

Walker, E., A. V. Hernandez and M. W. Kattan (2008). "Meta-analysis: Its strengths and limitations." Cleveland Clinic Journal of Medicine **75**(6): 431-439.

Wan, Y., M. Kertesz, R. C. Spitale, E. Segal and H. Y. Chang (2011). "Understanding the transcriptome through RNA structure." Nature Reviews Genetics **12**(9): 641-655.

Wang, J., K. R. Coombes, W. E. Highsmith, M. J. Keating and L. V. Abruzzo (2004). "Differences in gene expression between B-cell chronic lymphocytic leukemia and normal B cells: a meta-analysis of three microarray studies." Bioinformatics **20**(17): 3166-3178.

Wang, J. B., V. Nygaard, B. Smith-Sorensen, E. Hovig and O. Myklebost (2002). "MArray: analysing single, replicated or reversed microarray experiments." Bioinformatics **18**(8): 1139-1140.

Wang, Z. Y., S. Kusam, V. Munugalavadla, R. Kapur, R. R. Brutkiewicz and A. L. Dent (2006). "Regulation of Th2 cytokine expression in NKT cells: Unconventional use of Stat6, GATA-3, and NFAT2." Journal of Immunology **176**(2): 880-888.

Watford, W. T., B. D. Hissong, J. H. Bream, Y. Kanno, L. Muul and J. J. O'Shea (2004). "Signaling by IL-12 and IL-23 and the immunoregulatory roles of STAT4." Immunological Reviews **202**: 139-156.

Watterson, S., M. L. Guerriero, M. Blanc, A. Mazein, L. Loewe, K. A. Robertson, H. Gibbs, G. Shui, M. R. Wenk, J. Hillston and P. Ghazal (2013). "A model of flux regulation in the cholesterol biosynthesis pathway: Immune mediated graduated flux reduction versus statin-like led stepped flux reduction." Biochimie **95**(3): 613-621.

Watterson, S., M. L. Guerriero, M. Blanc, A. Mazein, L. Loewe, K. A. Robertson, H. Gibbs, G. H. Shui, M. R. Wenk, J. Hillston and P. Ghazal (2013). "A model of flux regulation in the cholesterol biosynthesis pathway: Immune mediated graduated flux reduction versus statin-like led stepped flux reduction." Biochimie **95**(3): 613-621.

Wei, C. M., J. N. Li and R. E. Bumgarner (2004). "Sample size for detecting differentially expressed genes in microarray experiments." Bmc Genomics **5**.

Welch, B. L. (1947). "The Generalization of Students Problem When Several Different Population Variances Are Involved." Biometrika **34**(1-2): 28-35.

Wilkinson, B. (1951). "A Statistical Consideration in Psychological Research." Psychological Bulletin **48**(2): 156-158.

Willingham, A. T. and T. R. Gingeras (2006). "TUF love for "junk" DNA." Cell **125**(7): 1215-1220.

Xiang, Q., X. H. Dai, Y. Y. Deng, C. S. He, J. Wang, J. H. Feng and Z. M. Dai (2008). "Missing value imputation for microarray gene expression data using histone acetylation information." Bmc Bioinformatics **9**.

Yang, I. V., W. W. Jiang, H. R. Rutledge, B. Lackford, L. A. Warg, L. De Arras, S. Alper, D. A. Schwartz and D. S. Pisetsky (2011). "Identification of novel innate immune genes by transcriptional profiling of macrophages stimulated with TLR ligands." Molecular Immunology **48**(15-16): 1886-1895.

Yang, Y. H. and T. Speed (2002). "Design issues for cDNA microarray experiments." Nature Reviews Genetics **3**(8): 579-588.

Yang, Y. H. and T. Speed (2002). "Design issues for cDNA microarray experiments." Nature Reviews Genetics **3**(8): 579-588.

Yuen, T., E. Wurmbach, R. L. Pfeffer, B. J. Ebersole and S. C. Sealfon (2002). "Accuracy and calibration of commercial oligonucleotide and custom cDNA microarrays." Nucleic Acids Research **30**(10).

Zaykin, D. V., L. A. Zhivotovsky, P. H. Westfall and B. S. Weir (2002). "Truncated product method for combining P-values." Genetic Epidemiology **22**(2): 170-185.